IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 17, Issue 5, Ver. III (Sep. – Oct. 2015), PP 60-64
www.iosrjournals.org

A Survey on Privacy Preserving Data Mining Techniques

Deepa Tiwari, Raj Gaurang Tiwari
Department of Computer Science and Engineering Shri Ramswaroop Memorial College of Engineering &Management / UPTU Lucknow, Uttar Pradesh, India

Abstract: *Data mining is the extraction of the important patterns or information from large amount of data, which is used for decision making in future work. But the process of data collection and data dissemination may cause the serious damage for any organization, however may causes the privacy concerns. Sensitive or personal information of individuals, industries and organization must be kept private before it is shared or published. Thus privacy preserving data mining has become an important issue to efficiently hide sensitive information. Many numbers of methods and techniques have been proposed for privacy preserving data mining for hiding sensitive information. In this paper we provides our own overview which has taken from previous paper on privacy preserving data mining.*

Keyword: *Privacy preserving data mining, Sanitized database, Greedy approach, Transaction insertion, Genetic algorithm, Restrictive Patterns.*

I. Introduction

Data mining is the process of discovering of knowledge from huge amount of data, databases or data warehousing and this knowledge is used for decision making, process control, information management and query processing, however it can also disclosure of sensitive information about any individuals organization etc.

In recent year with the rapid growth of development in internet, data storage and data processing technologies, privacy preserving data mining has been drawn increasing attention. In order to make publicly available system secure, not only focus on that private information but also to make secure that certain inference channels have been blocked as well. confidential information lead to privacy concerns if the provide information is misused. Confidential information includes income, medical history, address, credit card number, phone number, purchasing behavior etc. some shared information among companies can be extracted and analyzed by any third parties, which may not only decrease the benefits of companies but also cause threats to sensitive data. According to the definition, Privacy is the quality or condition of being secluded from the presence or view of others [1] . On relating privacy with data mining, privacy implies keep information about individual from being available to others [2] . 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. The main objective of PPDM is to develop algorithms and techniques by which original data is modified in some way that mining process cannot extract original confidential data. The procedure of transforming the source database into a new database that hides some sensitive patterns or rules is called the sanitization process. Thus therefore privacy preserving data mining has becoming an increasingly important field of research concerned with the privacy driven from personally identifiable information when considered for data mining. In this paper, we discuss different approaches and techniques in the field of privacy preserving data mining and analyses the representative methods for privacy preserving data mining, and points out their merits and demerits.

II. Privacy preserving data mining-

Privacy preserving [9] 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.

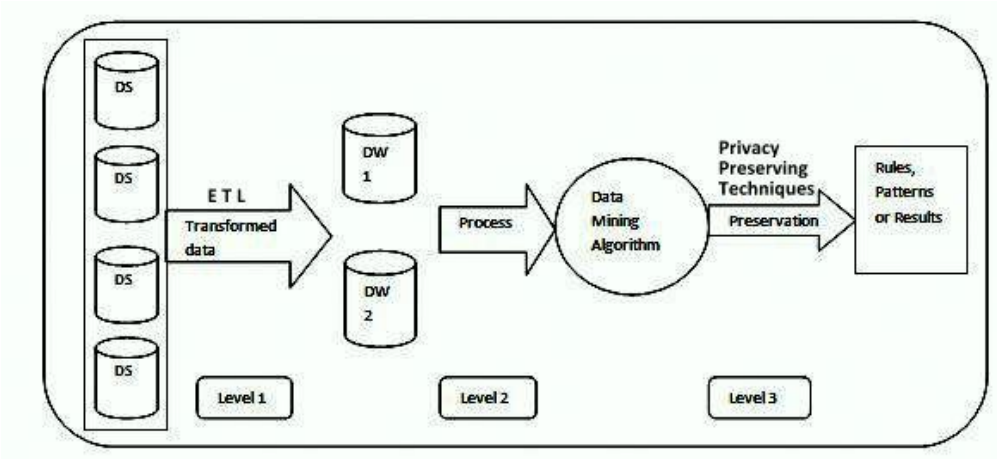

Fig. 1 Framework of privacy preserving data mining

III. Privacy preserving data mining techniques-

In this section we describe different privacy preserving data mining techniques which are developed like transaction insertion, transaction deletion, heuristic pattern, blocking based techniques etc.

IV. A greedy based approach for hiding sensitive itemsets by transaction insertion-

Chun-Wei Lin et.al. [3] proposed a greedy-based approach to hide sensitive itemsets by transaction insertion. The proposed approach first computes the maximal number of transactions to be inserted into the original database for totally hiding sensitive itemsets. The fake items of the transactions to be inserted are thus designed by the statistical approach, which can greatly reduce side effects in PPDM. the sensitive itemsets are then hidden by adding newly transactions into the original database, thus increasing the minimum count threshold to achieve the goal. It is, however, three factors should be taken as the consideration. First, the number of transactions should be serious determined for achieving the minimal side effects to totally hide the sensitive itemsets. In this part, sensitive itemsets are then respectively evaluated to find the maximal number of transactions to be inserted. Second, the length of each newly inserted transaction is then calculated according to the empirical rules in standard normal distribution. Last, the already existing large itemsets are then alternatively added into the newly inserted transactions according to the lengths of transactions which determined at the second procedure. This step is to avoid the missing failure of the large itemsets for reducing the side effects in the PPDM.

A greedy-based approach for data sanitization proposed three steps to insert new transactions into original database for hiding sensitive itemsets. In the first step, the safety bound for each sensitive itemset is then calculated to determine how many transactions should be inserted. Among the calculated safety bound of each sensitive itemset, the maximum operation is then used to get the maximal numbers of inserted transactions. Next, the lengths of inserted transactions are then evaluated through empirical rules in statistics as the standard normal distribution. In the third step, the count difference is then calculated between the sensitive itemsets and non sensitive frequent itemsets at each k-level (k-itemset). The non-sensitive frequent itemsets are then inserted into the transaction in descending order of their count difference. This property remains that the original frequent itemsets would be still frequent after the numbers of transactions are inserted for hiding sensitive itemsets. The above steps are then repeatedly progressed until all sensitive itemsets are hidden.

V. An Effective Heuristic Approach for Hiding Sensitive Patterns in Databases-

M.Mahendran et.al. [4] proposed a heuristic approach which ensures output privacy that prevent the mined patterns(itemsets) from malicious inference problems. An efficient algorithm named as Pattern-based Maxcover Algorithm is proposed. This algorithm minimizes the dissimilarity between the source and the released database; Moreover the patterns protected cannot be retrieved from the released database by an adversary or counterpart even with an arbitrarily low support threshold. we need to develop mechanisms that can lead to new privacy control systems to convert a given database into a new one in such a way to preserve the general rules mined from the original database. The procedure of transforming the source database into a new database that hides some sensitive patterns or rules is called the sanitization process[12]. To do so, a small number of transactions have to be modified by deleting one or more items from them or even adding noise to the data by turning some items from 0 to 1 in some transactions. The released database is called the sanitized database. On one hand, this approach slightly modifies some data, but this is perfectly acceptable in some real applications[13,14].

The optimal sanitization has been proved to be an NP-hard problem. To alleviate the complexity of the optimal sanitization, some heuristics could be used. A heuristic does not guarantee the optimal solution but usually finds a solution close to the best one in a faster response time.

Given the source database (D), and the restrictive patterns(RP), the goal of the sanitization process is to protect RP against the mining techniques used to disclose them. The sanitization process decreases the support values of restrictive patterns by removing items from sensitive transactions. This process mainly includes four sub-problems:

1. Identifying the set of sensitive transactions for each restrictive pattern;
2. Selecting the partial sensitive transactions to sanitize;
3. Identify the candidate item(victim item) to be removed;
4. Rewriting the modified database after removing the victim items.

VI. The GA-based algorithms for optimizing hiding sensitive itemsets through transaction deletion-

Chun-Wei Lin et.al.[5] proposed two algorithms, simple genetic algorithm to delete transaction (sGA2DT) and pre-large genetic algorithm to delete transaction (pGA2DT) based on the genetic algorithm [6]. Genetic algorithms (GAs) [6] are able to find optimal solutions using the principles of natural evolution. A novel GA-based framework consisting of two designed algorithms is proposed to conquer the optimal selective problems of heuristic approaches. A flexible evaluation function, consisting of three factors with their adjustable weightings, is designed to determine whether certain transactions are optimal to be deleted for the purpose of hiding sensitive itemsets. The proposed algorithms were intended to delete a pre-defined number of transactions for hiding sensitive itemsets. A simple genetic algorithm and pre-large concepts [7,8] are also considered to reduce the execution time for rescanning the original databases for chromosome evaluation, and the number of populations in the proposed algorithms. A straightforward approach (Greedy) is designed as a benchmark to evaluate the performance of the two proposed algorithms as a simple genetic algorithm to delete transactions (sGA2DT), and a pre-large genetic algorithm to delete transactions (pGA2DT) with regards to the execution time, the three side effects (hiding failures, missing itemsets, and artificial itemsets), and database dissimilarity in the experiments.

Genetic algorithm-

Holland[6] applied the Darwin theory of natural selection and survival of the fittest, into the field of dynamic algorithms and proposed the evolutionary computation of genetic algorithms (GAs) [6]. GAs are a class of search techniques designed to find a feasible solution in a limited amount of time. According to the principle of survival of the fittest, GAs generate the next population by various operations with each individual in the current population which represent a set of possible solutions. Three major operations used in GAs are described below:

1. **Crossover:** The offspring is generated from two chosen individuals in the population by swapping some attributes among the two individuals. The offspring inherits some characteristics from both of the two individuals (parents).
2. **Mutation:** Mutation: One or several attributes of an offspring may be randomly changed. The offspring may possess different characteristics from their parents. Mutation increases the possibility of achieving global optimization.
3. **Selection:** Excellent offspring are chosen for survival according to predefined rules. This operation keeps the population size within a fixed amount, and the good offspring have a higher possibility of getting into the next generation.

The first step to find the optimal solution of GAs is to define a chromosome to represent a possible solution. A chromosome is usually stated in a bit string. An initial population consisting of many chromosomes, also called individuals, is defined as a set of possible solutions. The three genetic operations (crossover, mutation, and selection) are then performed on chromosomes for the next generation. Each chromosome is evaluated by the designed fitness function to evaluate its goodness. This procedure is recursively performed until the termination criterion is satisfied.
The entire GA process is shown in Fig. 2

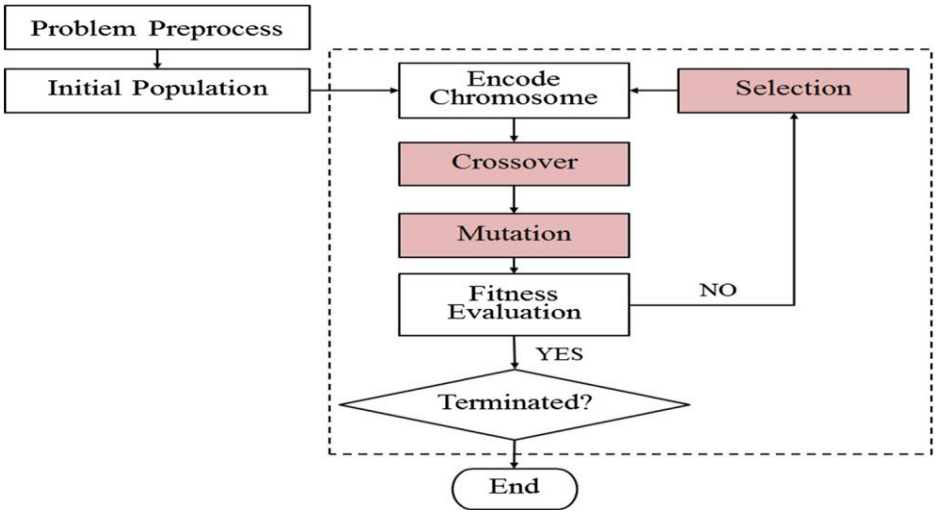

Fig. 2 Flowchart of Gas.

VII. Privacy preserving data mining based on the blocking based technique-

In blocking based technique [10][11], 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 [10] 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.

VIII. Conclusion-

In this paper we present a review of the privacy preserving data mining We discussed a variety of data modification techniques such as greedy based approach for hiding sensitive itemsets by transaction insertion, heuristic approach for hiding the heuristic patterns, genetic algorithm based techniques through transaction deletion and blocking based techniques. In this paper we also gives the overview of genetic algorithm. In all the techniques database is protected by privacy preservation on the basis of transaction insertion, transaction deletion, secure the frequent pattern by transaction deletion and replacement of the transaction. While all the purposed methods are only approximate to our goal of privacy preservation, we need to further perfect those approaches or develop some efficient methods.

Reference-

[1]. The free dictionary.Homepage on Privacy [Online]. Available: http://www.thefreedictionary.com/privacy.
[2]. M. Prakash, G. Singaravel, "A New Model for Privacy Preserving Sensitive Data Mining", in proceedings of ICCCNT Coimbatore, India, IEEE 2012.
[3]. Chun-Wei Lin, Tzung-Pei Hong, Chia-Ching Chang, and Shyue-Liang Wang "A Greedy-based Approach for Hiding Sensitive Itemsets by Transaction Insertion", Journal of Information Hiding and Multimedia Signal Processing, Volume 4, Number 4, October 2013.
[4]. M.Mahendran, Dr.R.Sugumar, K.Anbazhagan, R.Natarajan "An Efficient Algorithm for Privacy Preserving Data Mining Using Heuristic Approach", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 1, Issue 9, November 2012.
[5]. Chun-Wei Lin · Tzung-Pei Hong ·Kuo-Tung Yang ·Leon Shyue-LiangWang "The GA-based algorithms for optimizing hiding sensitive itemsets through transaction deletion", Published online: 20 September 2014, © Springer Science+Business Media New York 2014.
[6]. Holland JH (1992) Adaptation in natural and artificial systems. MIT Press.
[7]. Hong TP, Wang CY (2007) Maintenance of association rules using pre-large itemsets. In: Intelligent databases: technologies and applications, pp 44–60.
[8]. Hong TP, Wang CY, Tao YH (2001) A new incremental data mining algorithm using pre-large itemsets. Intell Data Anal 5:111–129

[9]. 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.
[10]. 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.
[11]. 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.
[12]. Gayatri Nayak and Swagatika Devi (2011), 'A Survey On Privacy Preserving Data Mining: Approaches And Techniques', International Journal of Engineering Science and Technology pp.2127-2133
[13]. Guang Li and Yadong Wang (2011), 'Privacy-Preserving Data Mining Based on Sample Selection and Singular Value Decomposition', Proceedings of the IEEE International Conference on Internet Computing and Information Services , pp.298-301
[14]. Jain Y.K. (2011), 'An Efficient Association Rule Hiding Algorithm for Privacy Preserving Data Mining', International Journal of Computer Science and Engineering, pp.96-104