International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 06, June-2015

A Novel Approach of Data Sanitization using Privacy Preserving Data Mining

Ashish Chouhan[1]
[1] M.Tech Scholar,
Department of IT,
University Institute of Technology, BU,
Bhopal, M.P. 462026,India

Dr. Poonam Sinha[2]
[2]Head,
Dept. of E.C.E and I.T.,
University Institute of Technology, BU,
Bhopal, M.P. 462026, India

Sunita Gond[3]
[3]Asst. Prof.,
Department of I.T.,
University Institute of Technology, BU,
Bhopal, M.P. 462026, India

***Abstract* - Privacy preserving data mining (PPDM) is a popular as well as interesting topic in the research community. The important issue is how to make a balance between privacy protection and knowledge discovery in the sharing process. One of the existing privacy preserving utility mining and two algorithms, HHUIF (Hiding High utility item First Algorithm) and MSICF (Maximum Sensitive ItemsetsConict First algorithm), to conceal the sensitive itemsets so that the antagonist cannot mine them from the modified database. The work also minimizes the impact on the sanitized database of hiding sensitive item sets. In order to address this sanitization we introduced a privacy preserving data mining using secure hash algorithm technique to modify itemset based on threshold value. We primarily focus on protecting privacy in database. By finding sensitive itemset we calculate SHA of these sensitive itemset and apply proposed algorithm to modify itemset. On different value of threshold we calculate value of hiding failure and miss cost. At last we summarized that as value of threshold increased value of hiding failure and missing cost decreased.**

I. INTRODUCTION

In the past few years Privacy Preserving Data Mining (PPDM) [6] is a relatively new research area in data mining. It aims to prevent the violation of privacy that might result from data mining operations on data sets [7, 9].PPDM algorithms modify original data sets so that privacy is preserved even after the mining process is activated, while minimally affecting the mining results quality.In 1996, Clifton et al. [10] analyzed that data mining can bring about threat against databases and addressed possible solutions to achieve privacy protection of data mining. In 2007, Podpecan et al. [4] proposed that utility based mining will play an important role. Utility mining is used to find out the high utility itemsets. User defined utility is based on the information not available in the transaction dataset. It often requires user preference and then it can be represented by an external utility table.

Some literary works based on privacy preserving utilitymining are discussed in the literature. Hence, this study focuses on privacy preserving data mining and presents novel algorithm Privacy Preserving Data Mining Using Secure Hash Algorithm(PPDMSHA), to achieve the privacy in the database(to achieve the goal of hiding sensitive itemsets), so the adversaries cannot extract them from the modified database. The process of converting theoriginal database into the sanitized one is called sanitization. The rest of this paper isorganized as follows. Section 2 reviews the related works. Section 3 proposed PPDMSHA algorithm. Section 4 discusses the experimental results and evaluates the performance ofthe proposed algorithm. Finally, Section 5 concludes the present work.

II. RELATED WORKS

Yeh, Hsu and Wen [1] have focused on privacy preserving utility mining andproposed two novel algorithms called HHUIF (Hiding High utility item First Algorithm) andMSICF (Maximum Sensitive ItemsetsConict First algorithm), in order to achieve thegoal of hiding sensitive itemsets, so that the adversaries cannot mine them from themodified database. On the other hand, they have also minimized the impact on the sanitized database of hiding sensitive itemsets. The experimental results have shown thatthe HHUIF achieved a lower miss cost than MSICF on two synthetic datasets. On theother hand, MSICF generally has a lower difference ratio between original and sanitizeddatabases than the HHUIF.

Rajalaxmi and Natarajan [3] have proposed on utility mining model. Data Sanitizationis the process to conceal the sensitive itemsets present in the source database with appropriate modifications and release the modified database. The problem of finding an optimum solution for the

sanitization process which minimizes the non-sensitive patterns lost is NP-hard. Several researches in data sanitization, this approach hide the sensitive itemsets by reducing the support of the itemsets which considers only the presence or absence of itemsets. However in real world scenario the transactions contain the purchased quantities of the items with their unit price. Hence it is essential to consider the utility of itemsets in the source database.In order to address this utility miningmodel was introduced to find high utility itemsets. Here, the utility of the itemsets and propose a novel approach for sanitization such that minimalchanges are made to the database with minimum number of non-sensitive itemsets removed from the database.

Li, Yeh and Chang [5] have proposed a MICF: An effective sanitization algorithm, in order to conceal restrictive itemsets (patterns) contained in the source database, a sanitization process transforms the source databaseinto a released database that the counterpart cannot extract sensitive rules from. The transformed result also conceals non-restrictiveinformation as an unwanted event, called a side effect or the ‘‘misses cost’’. The problem of finding an optimal sanitization method, which
conceals all restrictive itemsets but minimizes the misses cost, is NP-hard. To address this challenging problem, this study proposes themaximum item conflict first (MICF) algorithm. Theexperimental results have shown that the proposed method is effective, has a lowsanitization rate, and can generally achieve a significantly lower misses cost than those achieved by the MinFIA, MaxFIA, IGA andAlgo2b methods in several real and artificial datasets.

Oliveira and Zaine [8] have proposed a framework for enforcing privacy in mining frequent patterns. They combined, in a single framework, techniques for efficiently hiding restrictive patterns and a set of algorithms to sanitize a database. In order to address the privacy requirements in mining hidden pattern is to look for a balance between hiding restrictive patterns and disclosing non-restrictive ones.

III. PROPOSED METHODOLOGY

In privacy preserving data mining using sanitization based approach we develop a new algorithm called *“Privacy Preserving Data Mining Using SHA”* for achieving privacy in the database. There are following steps:

3.1 Privacy Preserving Data Mining Using SHA

1. Create a database DB which has large no. of data items.
2. Find out sensitive data items from these items based on utility mining threshold specific value.
3. Modify these sensitive values which have to modify.
4. Calculate MD5 (SHA) checksum of this utility itemset.
5. Remove all the digits from A to F from the hex number.
6. Subtract the first x left hand side digits from the modified hex.

Privacy preserving data mining using data sanitization approaches provide privacy to sensitive data item set.

IV. EXPERIMENTAL RESULTS

For simulating the results hiding sensitive data items using secure hash algorithm we use Apache web server, Php andMysql.

4.1 Data set:

We used the IBM synthetic data generator [11] to generate datasets. To check performance of the proposed algorithm for privacy preserving data mining using secure hash algorithm, we can evaluate it practically using a bank dataset containing 1000 data items respectively. In bank dataset we find out value of hiding failure and miss cost.

Experiment done on 1000 no. of data items and results shown on the threshold value containing 2000,3000, 4000, 5000, 6000, 7000, 8000 respectively and calculated value of hiding failure and miss cost respectively.

TABLE 1.shows calculated value of hiding failure and missing cost using PPDMSHA algorithm

Threshold value	No. of sensitive items found	PPDMSHA Hiding Failure(HF)	PPDMSHA Missing cost(MC)
2000	998	4.99	8.98
3000	997	4.985	8.97
4000	996	4.98	8.96
5000	995	4.975	8.95
6000	995	4.975	8.95
7000	994	4.97	8.94
8000	992	4.96	8.92
9000	991	4.955	8.91
10000	990	4.95	8.9
50000	955	4.775	8.55
100000	891	4.455	7.91
500000	525	2.625	4.25

TABLE 2.shows comparison of missing cost of previous algorithm and using PPDMSHA algorithm[2]

Threshold value	HHUIF algorithm Missing cost (MC)	PPDMSHA algorithm Missing cost(MC)
2000	68.04	8.98
3000	62.04	8.97
4000	50.0	8.96
5000	35.71	8.95
6000	39.38	8.95
7000	46.74	8.94
8000	32.61	8.92

4.2Performance Analysis:

Our proposed PPDMSHA algorithm performance is compared with the HHUIF algorithm given in [2]. The performance analysis is carried out by the threshold value as 2000, 3000, 4000, 5000, 6000, 7000 and 8000. The performance measures of our proposed and conventional algorithms are shown in the following table 2.

The performance measures are described below,

(a) *Miss Cost (MC):* the ratio of valid itemsets presented in the original database and sanitized database. The miss cost is measured as follows:

$$MC = \frac{|U(D) - U(D')|}{|U(D)|} \quad (1)$$

where U(D) and U(D') denote the sensitive itemsets discovered from the original database D and the sanitized database D' respectively.

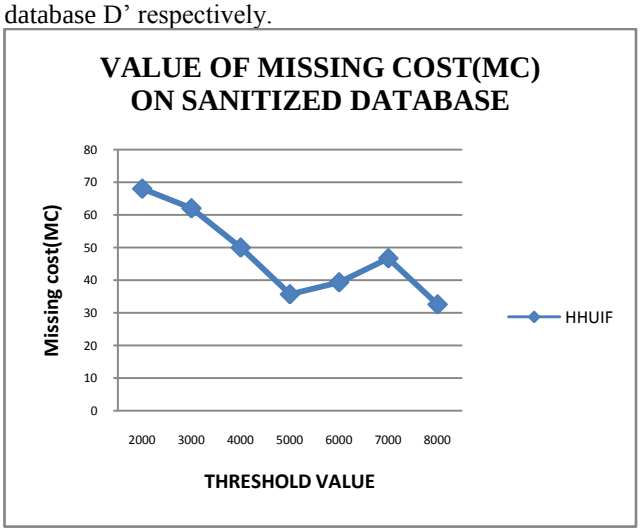

Figure 1.shows value of missing cost(mc) on different threshold value using HHUIF algorithm

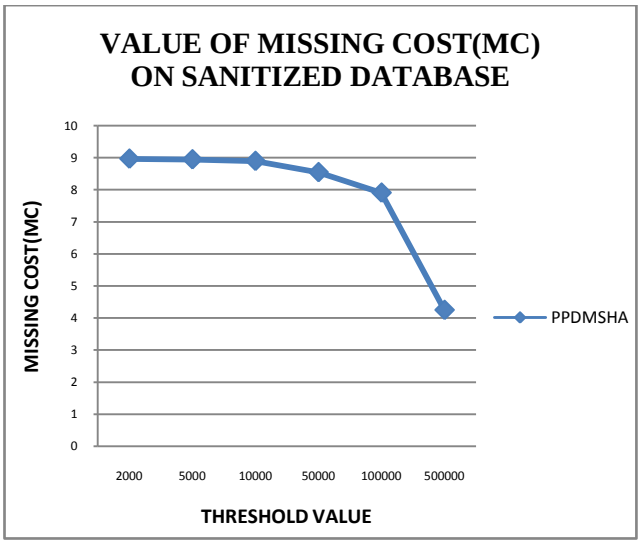

Figure 2.shows value of missing cost (mc) on different threshold value using PPDMSHA algorithm

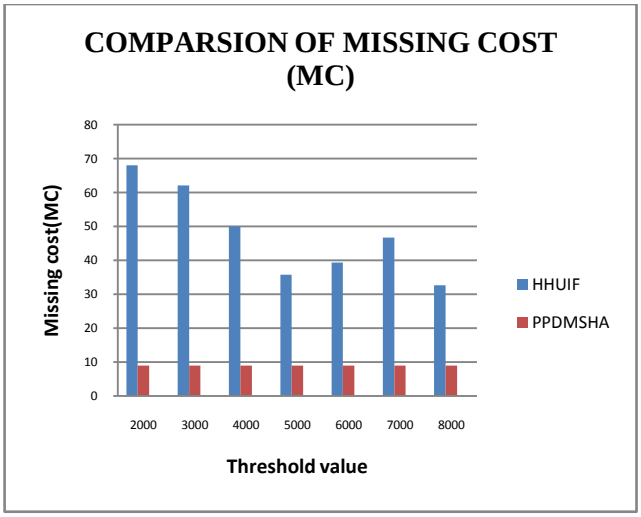

Figure 3.shows comparison of missing cost (mc) using PPDMSHA algorithm and HHUIF algorithm

Figure 1, 2 and 3 illustrate the performance of HHUIF and PPDMSHA algorithms in different threshold values with different performance measures. The lower miss cost value shows that our sanitization database contains more valid items than the original database.

V. CONCLUSION

In this study, we present Data sanitization utilizing secure hash algorithm to reduce the impact on the source database for the privacy preserving data mining. This algorithm is predicated on modifying the database containing the sensitive itemsets so that the utility value can be reduced below MinUtility threshold value. There is no possible way to reconstruct the pristine database from the Sanitized one. In our experimental results, PPDMSHA has the lower miss costs in datasets.

REFERENCES

[1] J. S. Yeh, P. C. Hsu and M. H. Wen, "HHUIF and MSICF: Novel algorithms for privacy preserving utility mining",*Department of Computer Science and Information Management, Providence University*, 200 Chung Chi Rd., Taichung 43301, Taiwan 2011.

[2] Bi Ru Dai, "Hiding Frequent Patterns in the Updated Database", *Information Science and Applications (ICISA), International Conference, Pages 1-8,* 2010.

[3] R. R. Rajalaxmi and A. M. Natarajan, "A Novel Sanitization Approach for Privacy Preserving Utility Itemset Mining",*Computer and Information Science,* Vol. 1, No. 3, August 2008.

[4] Vid Podpecan, Nada Lavrac and Igor Kononenko, "A Fast Algorithm for Mining Utility Frequent Itemsets",*Jozef Stefan Institute, Ljubljana, Slovenia University of Nova Gorica, Nova Gorica, Slovenia University of Ljubljana, Faculty of Computer and Information Science, Ljubljana, Slovenia,* 2007.

[5] Yu Chiang Li, Jieh Shan Yeh and Chin Chen Chang, "MICF: An effective sanitization algorithm for hiding sensitive patterns on data mining", *Advanced Engineering Informatics*, 21 (2007), 269-280.

[6] V.Verykios, E.Bertino, I.G. Fovino, L. P. Provenza, Y. Saygin and Y. Theodoridis, "State of the art in privacy preserving data mining", *SIGMOD Record*, Vol. 33, No. 1, PP. 50-57, March 2004.

[7] Chris Clifton, Murat Kantarcioglu andJaideepVaidya,"Defining Privacy for Data Mining", ,*Department of Computer Sciences Purdue University*, PP. 191-207, 2004.

[8] S. R. M. Oliveira and O. R. Zaine, "A framework for enforcing privacy in mining frequent patterns", *Technical Report, TR02-13, Computer Science Department, University of Alberta*, Canada, June 2000.

[9] Yehuda Lindell and Benny Pinkas, "Privacy Preserving Data Mining", in: *Bellare, M. eds.(2000) Advances in Cryptology CRYPTO 2000 Springer, Heidelberg,* PP. 36-54.

[10] C. Clifton and D. Marks, "Security and privacy implications of data mining", in *Proceedings of the 1996 ACM SIGMOD Workshop on Data Mining and Knowledge Discovery*, PP. 15-19, 1996.

[11] IBM Almaden Research Center. Synthetic data generation code for associations and sequential patterns. http://www.almaden.ibm.com/cs/projects/iis/hdb/Projects/data_mining/mining.shtml