ISSN No. 0976-5697

Volume 6, No. 8, Nov-Dec 2015

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Analysis of Methodologies for Hiding Sensitive Frequent Itemsets using Border Based Approach

Aarthna Maheshwari
Research Scholar: Information Technology
TechnologMahakal Institute Of Technology
Ujjain, India

Khushboo Pawar
Associate Professor: Information
Mahakal Institute Of Technology
Ujjain, India

Abstract:Data mining is a technique that blends traditional data analysis methods with modern sophisticated algorithms for processing large amount of data. However, misuse of this technique may lead to disclosure of sensitive information. Privacy preserving data mining is the new effort in the screening of sensitive information. Association rule hiding methodologies aim at sanitising the original database in a way that now rule can be mined from the original database. This paper contains a survey of various sensitive information hiding methodologies.

Keywords: Data Mining, PPDM, Support , Confidence, Association rule hiding, sensitive information hiding.

I. INTRODUCTION

Data mining [1] is a process of automatically discovering useful information in large data repositories. Data mining techniques are deployed to scour large databases in order to find novel and useful patterns that might otherwise remain unknown. They also provide capabilities to predict the outcome of a future observation. Example: Whether a customer will buy cheese with bread or not. Data mining techniques can be used to support a wide range of business intelligence applications such as customer profiling, fraud detection. It can also help companies answer various questions like "What products should they use in new scheme for cross selling for profitability"
Association rule hiding [6] methodologies are used for the following purposes: No association rule is considered as sensitive from the owner's perspective and can be mined from the original database at pre-specified thresholds of confidence and support, can also be revealed from the sanitized database, when the database is mined at the same or at higher thresholds. All the non-sensitive association rules that appear when mining the database at pre-specified thresholds of confidence and support can be successfully mined from sanitized database at the same thresholds or higher and no association rule that was not derived from the original database at the pre-specified thresholds can be derived from the new database at the same or higher level.

II. BACKGROUND

A company has an original database (D) consisting of all the details of the transactions and company's information. It can be gestated as follows:

Let T be a set of transactions in D, uniquely identified by a transaction number Tid, consisting of a set of items I={i1,i2,i3,i4….}. An association rule is an implication of the form X → Y iff X is the subset of T, y is the subset of T and the intersection of X and Y is null. The association rule mining works as follows:

We find all the itemsets that appear frequently called frequent itemsets, so as to be considered relevant and derive them from association rules that are strong enough to be considered interesting. We aim at presenting some of these rules called sensitive frequent itemsets from being disclosed. The classification of these itemsets into sensitive and non-sensitive is based on company's privacy policy or because the disclosure of certain items may incur loss to the company.

The strength of an association rule is measured based on two parameters: support and confidence.

a) Support, $s(X \rightarrow Y) = \frac{| X \cup Y |}{N}$

Where: N is the total number of transactions in a database.

$|X \cup Y|$ is the support count of $X \cup Y$

b) Confidence, $c(X \rightarrow Y) = \frac{| X \cup Y |}{|X|}$

Where $|X \cup Y|$ is the support of $X \cup Y$ and $|X|$ is the support of itemset X.

The itemset is called frequent if and only if support is greater than or equal to minimum support value called threshold support.

To hide the sensitive frequent itemsets we have two approaches:

A) Decrease the support of the itemsets below the minimum threshold called minimum support threshold (MST).
B) Decrease the confidence of the itemsets below the minimum threshold called minimum confidence threshold (MCT).

© 2015-19, IJARCS All Rights Reserved

After hiding the sensitive frequent itemsets we get a new database called sanitized database (D') from which the sensitive frequent itemsets cannot be mined.

III. TECHNIQUES

We can classify the privacy preserving approaches into the following categories:

A. HEURISTIC TECHNIQUES

Heuristic techniques are used to solve the problem of selective data modification since selective data sanitization is an NP-Hard problem. Heuristic techniques are used to speed up the process of database sanitization. It was first proposed by M. Atallah, E. Bertino, A. Elmagarmid, M. Ibrahim, and V. S. Verykios[10].

- ***Blocking technique:***

In blocking technique [7] [13] [15], we change the known-values to "?" or to a known value. Now since the exact value is not known so support and confidence cannot be precisely defined and hence we have a Minimum support range and Minimum confidence range called Universal safety margin and if the support and confidence of the sanitized database are in the range then the security may not be breached.

But there are some problems associated with this approach:

I does not hide all the rules containing hidden itemsets say if X→Y and X→Z are two rules then it may hide X→Z but not X→Y i.e. it will not provide an optimal solution. Number of scans required in saygen's approach are higher than other algorithms because no of database scans are determined by the large itemsets and partial support transaction. Therefore it deteriorates the time performance of the algorithm. It also decreases the quality of the sanitized database by the addition of ghost noise.

- ***Distortion Scheme:***

In this scheme [3] [11] [14], we convert the database from D to D' in such a manner that the sensitive itemsets cannot be mined from the sanitized database. In this we change a set of 1's to 0's or vice versa in such a way that the support of the sensitive itemset is reduced.

The work in [2] showcases that sanitization of database using association rules as well as itemsets taking into account some assumptions. However it does not track the impact of modification of the items on the original database.

The paper [4] [8] tries to strike a balance between the privacy of data and the accuracy of the mining process. It tries to curb the fact that a non-sensitive information will generate a sensitive information. It scans the database multiple times. It is an item-restriction based algorithm in two aspects. First it involves deletion of items and it also it hides non sensitive items. It is fairly good for those datasets with a lower support range as higher support range will have a higher impact on the database.

According to [5] a balance can be stroked between data privacy and data mining if we use a matrix based approach. It removes the possibility of a database from suffering from forward inference attack and hiding failure.

The paper [19] extends the above matrix approach by using orthogonal matrix in the context of computing inner product and Eucledian distance and then using the perturbed version of the database. Even if the attacker gets access to perturbed database he may only be able to find approximate value of the perturbed database.

B. BORDER BASED APPROACH

It uses condense representation of itemsets (borders) to distinguish between frequent and infrequent itemsets. It is an item restriction based approach that uses greedy approach to select the transaction which will have minimum side effect.

The paper [16] highlights the importance of including the relative frequency of the itemsets in order to promote better results when a different support and confidence values are used. It is based on the concept of border revision. This algorithm suffers from serious performance (time) issues since it continuously monitors the itemsets and their supports. Also it is a less effective method as compared to other methods of the same approach. Also in this particular algorithm border representation is lossy and is underutilized since the Bd+ items once generated will be used as a close approximation every time and hence best results are not produced.

The extension of the previous work [9] uses the same approach used in zero-sum game (decision theory) for minimizing the distance possible distance to minimize the impact on the database. It uses the concept of border-revision. This approach has better performance if the itemsets are closely coupled. And its performance degrades slightly if the number of items increases

C. EXACT APPROACH

This approach is use to get optimal or suboptimal solution as the methods explained above use heuristics may get struck on local minima.[17] the approach uses border revision approach [12] and aims to clearly delineate the status of each itemsets based on certain constraints. At each stage, the constraints propagates to an itemsets to enforce local consistency for simplifying the problem.

The paper [18] restricts the sanitization of database through the hiding of itemsets rather than the rules. It assumes that the sanitized and original database should have minimum difference. It does not allow the production of rules having sanitized itemsets. It has the same size as that of original database.

IV. CONCLUSION:

We have provided a classification of various privacy preserving algorithms.. This paper adds to the demand of ever-increasing interest of researcher in this field. At this

© 2015-19, IJARCS All Rights Reserved

stage many algorithms have been developed but many of them suffer from time complexity, accuracy and privacy issues. Therefore, there is much scope for development of privacy preserving methodologies.

V. REFERENCES

[1] R. Agarwal, R. Srikant, "Fast algorithms for mining association rules," In: Proc. 20th Int'l Conf. Very Large Databases, 1994, pp. 487-499.

[2] Elena Dasseni, Vassilios S. Verykios, Ahmed K.Elmagarmid, and Elisa Bertino, "Hiding Association Rules by using Confidence and Support," In Proceedings of the 4th Information Hiding Workshop (2001), pp.369– 383

[3] V. S. Verykios, A. K. Emagarmid, E. Bertino, Y. Saygin, and E. Dasseni. Association rule hiding. IEEE Transactions on Knowledge and Data Engineering, 16(4):434–447, 2004.

[4] Stanley R. M. Oliveira and Osmar R. Zaiane, "Privacy preserving frequent itemset mining, In Proceedings of the IEEE ICDM Workshop on Privacy, Security and Data Mining (2002), pp.43–54

[5] E.T. Wang, G. Lee, Y.T. Lin, "A novel method for protecting sensitive knowledge in association rules mining," In: Proceedings of the 29th IEEE Annual International Computer Software and Applications Conference (COMPSAC'05), Edinburgh, Scotland, 2005, pp.511–516

[6] Rakesh Agrawal and Ramakrishnan Srikant, "Privacy-preserving data mining," In Proceedings of the ACM SIGMOD Conference on Management of Data (2000), pp.439–450

[7] Y. Saygin, V.S. Verykios, and C. Clifton, "Using Unknowns to Prevent Discovery of Association Rules," ACM SIGMOD Record, vol. 30, no. 4, pp. 45-54, 2001.

[8] S.R.M. Oliveira and O.R. Zaı ¨ne, "Privacy Preserving Frequent Itemset Mining," Proc. IEEE ICDM Workshop Privacy, Security, and Data Mining, pp. 43-54, 2002.

[9] Moustakides G, Verykios VS, A max-min approach for hiding frequent itemsets. In: Proceedings of 6th IEEE International Conference on Data Mining, pp 502-506.

[10] M. Atallah, E. Bertino, A. Elmagarmid, M. Ibrahim, and V. S. Verykios. Disclosure limitation of sensitive rules. In Proceedings of the 1999 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX), pages 45–52, 1999

[11] A. Amiri. Dare to share: Protecting sensitive knowledge with data sanitization. Decision Support Systems, 43(1):181–191, 2007.

[12] H. Mannila and H. Toivonen. Levelwise search and borders of theories in knowledge discovery. Data Mining and Knowledge Discovery, 1(3):241–258, 1997.

[13] Y. Saygin, V. S. Verykios, and C. W. Clifton. Using unknowns to prevent discovery of association rules. ACM SIGMOD Record, 30(4):45–54,2001.

[14]Y.Saygin,V.S.Verykios,and A. K. Elmagarmid.Privacypreservingassociationrulemining. In Proceedings of the 2002 International Workshop on Research Issues in Data Engineering: Engineering E–Commerce/E–Business Systems (RIDE), pages 151–163, 2002.

[15] E. Pontikakis, Y. Theodoridis, A. Tsitsonis, L. Chang, and V. S. Verykios. A quantitative and qualitative analysis of blocking in association rule hiding. In Proceedings of the 2004 ACM Workshop on Privacy in the Electronic Society (WPES), pages 29–30, 2004.

[16] X.SunandP.S.Yu. Hidingsensitivefrequentitemsetsbyaborder–basedapproach. Computing science and engineering, 1(1):74–94, 2007.

[17] Gkoulalas-Divanis A, Verykios VS (2009) Hiding sensitive knowledge without side effects. Knowl Inf Syst 20(3): 263–299.

[18]A. Gkoulalas-Divanis and V. S. Verykios. An integer programming approach for frequent itemsethiding. InProceedingsofthe15thACMInternationalConferenceonInformationand Knowledge Management (CIKM), pages 748–757, 2006.

[19] Liu, K., Kargupta, H., Ryan, J.: Random Projection-Based Multiplicative Data Perturbation for Privacy Preserving Distributed Data Mining. IEEE Transactions on Knowledge and Data Engineering 18(1), 92–106 (2006).

© 2015-19, IJARCS All Rights Reserved