ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 5, Issue 11, November 2016

Data Sanitization for Preserving Data Privacy

S. Devika M.Sc,
M. Phil Research Scholar,
Department of Computer Science,
Government Arts College, Coimbatore.

Dr. D. Devakumari MCA., M.Phil., Ph.D.
Assistant Professor,
PG and Research Dept of Information Technology,
Government Arts College, Coimbatore.

Abstract

Data Mining has emerged as a very popular tool for extracting hidden knowledge from collection of huge amount of data. Major challenges of data mining are to find the hidden knowledge in the data while the sensitive information is not revealed. Many industry, defence, public sector and organization are facing risk or having security issues while sharing their data. Association rule hiding is one of the PPDM(Privacy Preserving Data Mining) technique to protect the sensitive association rules generated by association rule mining. This paper adopts heuristic approach for hiding sensitive association rules. The proposed technique makes the representative rules and hides the sensitive rules. The proposed algorithm to hide sensitive patterns in the form of frequent item sets and infrequent item sets from the database while controlling the impact of sanitization on the data utility using estimation of impact factor of each modification on non-sensitive item sets.

Keywords: Data Mining, Association Rule Hiding, Data Sanitization, Privacy Preserving Data Mining

Introduction

Data mining is the knowledge discovery process of finding the useful information and patterns out of large database. In recent times data mining has gained immense importance as it paves way for the management to obtain hidden information and use them in decision-making [14] Privacy preserving data mining (PPDM) come up with the idea of protecting sensitive data or knowledge to conserve privacy while data mining techniques can still be applied efficiently [1]. There have been two types of privacy concerning data mining [14], [8]: (1) data privacy, and (2) information privacy. In data privacy, the database is modified in order to protect sensitive data of individuals. Whereas in information privacy (e.g. clustering or association rule), the modification is done to protect sensitive knowledge that can be mined from the database. In other words data privacy is related to input privacy while information privacy is related to output privacy.

Privacy preserving data mining is a major research area for protecting sensitive data or knowledge. Association rule hiding

is one of the privacy preserving techniques to hide sensitive association rules. The main aim of all association rule hiding algorithm is to minimally modify the original database and see that no sensitive association rule is derived from it.

It protects sensitive information by providing sanitized database of original database on the internet or a process is used in such a way that private data and private knowledge remain private even after the mining process. It is PPDM due to which the benefits of data mining be enjoyed, without compromising the privacy of concerned individuals. PPDM Techniques can be classified over five dimensions [9].The first dimension is related to distribution of data i.e. Centralized or Distributed. The second dimension refers to the modification of original values of data that are to be released for data mining task. Modification is carried out using perturbation, blocking, aggregation, merging, swapping or sampling or any combination of these. The third dimension is that of data mining algorithms. The data mining algorithms applied on the transformed data to get useful nuggets of information that were hidden previously. The fourth dimension refers to whether the raw data or aggregated data should be hidden. The fifth and the final dimension refer to the techniques that are used for protecting privacy. Based on these dimensions, different PPDM techniques may be classified into following five categories [9][4].

Association rules are usually required to satisfy a user- specified minimum support and a user-specified minimum confidence at the same time. Association rule generation is usually split up into two separate steps:

1. First, minimum support is applied to find all frequent item sets in a database.

2. Second, these frequent item sets and the minimum confidence constraint are used to form rules.

There are three types of Association rule hiding algorithms namely border-based approaches, exact approaches and heuristic approaches. Heuristic approaches are very efficient and fast algorithms that modify the selected transactions from the database for hiding the sensitive knowledge [10].

Problem Definition

The objective of privacy preserving data mining is to hide certain confidential data so that they cannot be discovered

through data mining techniques. The existing algorithm used to hide the sensitive items using HIF(High Impact Factor)algorithm. It calculates sensitive and non sensitive items using only frequent itemsets.

The ARM (Association Rule Mining)algorithms include level wise algorithms [1] such as Apriori, and pattern- growth methods [9] such as FP-Tree and FP- Growth. The process of ARM contains two phases. In the first phase, the frequent item sets that satisfy α m(minimum Support) are generated and then in second phase, the association rules that satisfy βmin(minimum confidence) are derived from the frequent item sets.

Proposed Algorithm

The Proposed Work considers Reconstruction-Based Association Rule Hiding. First, Using the Apriori algorithm to find frequent item sets. and hiding the sensitive items using sanitization algorithm. The proposed work has to hide frequent and infrequent sensitive items. Both items are evaluated in market analysis. In the case of other datasets may be infrequent items will be sensitive. The infrequent items are to be hidden using MApriori Algorithm. In this algorithm, calculate the maximum and closed items and support threshold will auto generated.

Fig (1) Calculates the Sensitive and non sensitive items using frequent patterns.

Input:
Frequent item sets(FI),Victim item(VI),hidden item(HI)
Output:
Sensitive Itemsets(SI),Non sensitive itemsets(NSI)
Algorithm:
Step1: Indexing the Frequent item set
Step2: Fix a victim item(VI) that has higher support
Step3: Set the hidden items(HI) that has average support
Step4: Find a VI based on priority list
Step5: Then HI based on priority is SI

Fig (1)Pseudo Code for calculting iSensitive items &non sensitive items.

Sanitization Algorithm

It has four Major steps:

Step 1: Identify sensitive transactions for each restrictive association rule:

Step 2: For each restrictive association rule, identify a candidate item that shou be eliminated from the sensitive transactions.This candidate item is called Victin Item(VI).

Step 3:Based on the disclosure thredhold¥, calculate for each restrictive association rule the nuber of sensitive transactions that should be sanitized.

Step 4: Based on the number found in step 3, identify for each restrictive association rule the sensitive transactions that have to be sanitized and remove the victim item from them.

Illustrative Example

Let us consider a sample transactional database and extracted frequent items with minimum support threshold 20%

TID	ITEMS
1	A,B,E
2	A
3	A,D,C
4	C,E
5	B,C
6	D,A
7	C,A
8	E
9	A,E
10	B,E

Table (1) Sample Transaction Data

ITEMS	SUPPORT
A	60%
B	30%
C	40%
D	20%
E	50%
AC	20%
AD	20%
AE	20%
BE	20%

Table (2) Support values for each item.

Rules	Confidence
A E	33.33%

A D	33.33%
A C	33.33%
B E	66.67%
C A	50.00%
D A	100.00%
E A	40.00%
E B	40.00%

Table (3) Confidence values for each item.

In this transaction data encounters Victim item is A.

Sensitive	Non sensitive
A	CD
E	AD
C	AE
B	BE

Table(4) Sensitive & Non- Sensitive items

TID	ITEMS
1	*,B,E
2	*
3	*,D,C
4	C,E
5	B,C
6	D,*
7	C,*
8	E
9	*,E
10	B,E

Table (5) Sanitized Data

Infrequent Items

In infrequent items the support threshold are auto Generated.and maximal and closed items are calculated reverse process executed.

In this transaction data encounters Victim Value is D.

ITEMS	SUPPORT
A	60%
B	30%
C	40%
D	20%
E	50%
AC	20%
AD	20%
AE	20%
BE	20%

Table (6) Support values for each item

.

Rules	Confidence
A E	33.33%
A D	33.33%
A C	33.33%
B E	66.67%
C A	50.00%
D A	100.00%
E A	40.00%
E B	40.00%

Table (7) Confidence for each item

Sensitive	Non sensitive
D	A
CE	E
AB	C
BC	B

Table (8) Sensitive non Sensitive items

TID	ITEMS
1	A,B,E
2	A
3	A,*,C
4	C,E
5	B,C
6	*,A
7	C,A
8	E
9	A,E
10	B,E

Table (9) Sanitized data

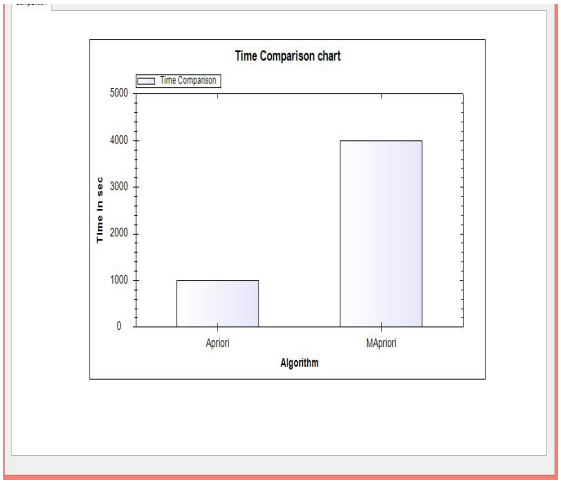


Fig (2) Time comparision for Apriori and Maporiori

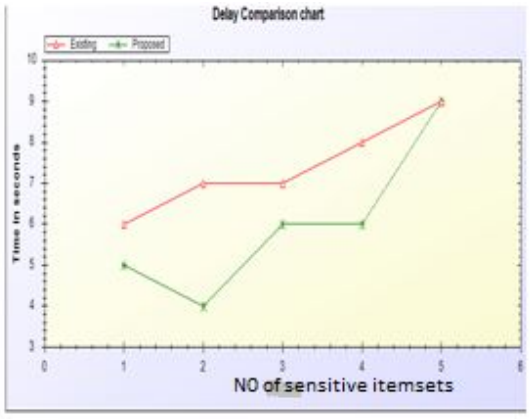
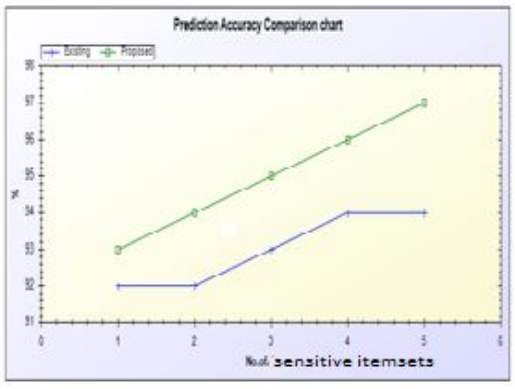


Fig (3) Delay Comparison for Sensitive items.



Fig(4) Prediction Accuracy for Sensitive items

Experimental Results

This Paperscompares the HIF with frequent items and infrequent senisitive items to hide using sanitization algorithm. the senisitive items are auto generated in the reconstruction based association rule hiding that has less computational time using HIF.

Conclusion

Both frequent and infrequent items are evaluated in market analysis. It may disclose patterns and various kinds of sensitive knowledge that are difficult to find otherwise, it may pose a threat to the privacy of discovered confidential information. Such information is to be protected against unauthorized access. Objective of this work is to propose a new strategy to avoid extraction of any kind of sensitive data. Data should be manipulated /distorted in such a way that sensitive information cannot be discovered through data mining techniques. This work discusses the threats to database privacy and security caused due to rapid growth of data mining and similar processes.

References

[1] A. Telikani1*, A. Shahbahrami2 and R. Tavoli3 (2015) "Data sanitization in association rule mining based on impact factor" Journal of AI and Data Mining Vol 3, No 2, 2015, 131-140.

[2] Amiri, A. (2007). Dare to share: "Protecting sensitive knowledge with data sanitization. Decision Support Systems", vol. 43, no. 1, pp. 181-191.

[3] Arpit Agrawal1 Asst. Professor Department of Computer Engineering Institute of Engineering & Technology Devi Ahilya University M.P., India Jitendra Soni Asst. Devi Ahilya "Secure Frequent Item set Hiding Techniques in Data Mining" International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3, Issue 2, February 2014.

[4] Bertino, E., Fovino, I. N. & Provenza, L. A. (2005)." A framework for evaluating privacy preserving data mining algorithms". Data Mining and Knowledge Discovery, vol. 11, no. 2, pp. 121-154.

[5] Bhavani Thuraisingham, The MITRE Corporation, USA "Privacy-Preserving Data Mining: Developments and Directions" IDEA GROUP PUBLISHING

[6] C. Clifton, "Using sample size to limit exposure todata mining" Journal of Computer Security, 2000.

[7] Dharmendra Thakur 1, Hitesh Gupta 2 " Study of Privacy Preserving Association Rule Mining Techniques" International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 11, November 2013.

[8]DhyanendraJain1,Amitsinhal2 ,Neetesh Gupta3,PriushaNarwariya4,Deepika Saraswat5,Amit Pandey6Technocrat Institute of Technology and Management, Gwalior, (M.P.), India "Hiding Sensitive Association Rules without Altering the Support of Sensitive Item(s")International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, No.2, March 2012.

[9] F. A. El-Mouadib and A. O. El-Majressi, "A study of multilevel association rule mining," in Proc. the Int. Arab Conf. Information Technology, Libya, Dec. 2010, pp. 14-16.

[10] Kasthuri S1 and Meyyappan T2 "HIDING SENSITIVE ASSOCIATION RULE USING HEURISTIC APPROACH" International Journal of Data Mining & Knowledge Management

Process (IJDKP) Vol.3, No.1, January 2013.

[11] Komal Shah, Amit Thakkar, Amit Ganatra," A Study on Association Rule Hiding Approaches" International Journal of Engineering and Advanced Technology (IJEAT), February 2012.

[12] N. Ravikumar "A Survey On Association Rule Mining" (IJARCCE) Volume 3, Issue- 1, Jan 2014.

[13] R.Natarajan1, Dr.R.Sugumar2, M.Mahendran3, K.Anbazhagan4 "Design and Implement an Association Rule hiding Algorithm for Privacy Preserving Data Mining" International Journal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 7, September 2012

[14] Saad M. Darwish, Magda M. Madbouly, and Mohamed A. El-Hakeem "A Database Sanitizing Algorithm for Hiding Sensitive Multi-Level Association Rule Mining" International Journal of Computer and Communication Engineering, Vol. 3, No. 4, July 2014

[15] Sarra Gacem1,Djamila,Mokeddem2 and Hafida Belbachir3 "Privacy Preserving Data Mining: Case of association rules" IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 3, No 1,

[16] S. Sangeetha 1,R. Kiruba 2 "A Hybrid Approach to Association Rule Hiding "International Journal of Computer Applications (0975 – 8887) National Conference on Information Processing and Remote Computing, NCIPRC 2015.

[17] Umesh Kumar Sahu 1, Anju Singh 2 "Approaches for Privacy Preserving Data Mining by Various Associations Rule Hiding Algorithms – A Survey".

[18] Wang P, "Survey on Privacy preserving data mining", International Journal of Digital Content Technology and its Applications, Vol. 4, No. 9, 2010.

[19] Wang P, "Survey on Privacy preserving data mining", International Journal of Digital Content Technology and its Applications, Vol. 4, No. 9, 2010.

[20] Yogendra Kumar Jain 1, Vinod Kumar Yadav 2, Geetika S. Panday 3 "An Efficient Association Rule Hiding Algorithm for Privacy Preserving Data Mining " International Journal on Computer Science and Engineering (IJCSE). May 2013 ISSN (Print): 1694-0814 | ISSN (Online): 1694-0784 .

.