ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

A Survey on Various Methodologies of Hiding Association Rules for Privacy Preserving

Divya Girdharlal Khatri *, Prof. Devangini Dave, Prof. Kishori Shekokar
Student, Department of Computer *Engineering, SIET, Vadodara, Gujarat India*
Professor, Department of Computer Engineering, SIET, Vadodara, Gujarat India
HOD, Department of Computer Engineering, SIET, Vadodara, Gujarat India

Abstract

Data mining is the useful technology to extract information or knowledge from large database. However, misuse of this technology may lead to the disclosure of sensitive information. Privacy preserving data mining (PPDM) is new research direction for disclosure of sensitive knowledge. There are various techniques used in PPDM to hide association rules generated by association rule generation algorithms. Main goal of privacy preserving data mining is to find association rules and to hide sensitive association rules. Association rule hiding is the process of modifying original database in such way that sensitive rules are disappeared. In this paper, a survey of various approaches of association rule hiding has been described.

Keywords: Data Mining, Privacy Preserving Data Mining, Association Rules, Sensitive Association Rule Hiding

Introduction

Data mining techniques enable people to find useful information from the large database. There are various techniques of data mining that are useful for extracting information from large database. Association rule is one of the most popular data mining techniques [1] in use. Many organizations shares there database for mutual benefits, this has increased the disclosure risks when the data is released to outside parties. For example, let a cloth store that purchase shirts from two companies, Levi's and United Colors of Benetton, Levi's applies data mining techniques and mines association rules related to United colors of Benetton, by applying data mining techniques to sensitive association rules of United colors of Benetton, Levi's had found that United colors of Benetton is offering 40% of discount and also customer's buy jeans with shirt, so Levi's offers 50% discount on shirts and also 20% of discount on jeans, This is how customers of United Colors of Benetton will now move to Levi's. This Scenario leads to the research of sensitive knowledge (or rule) hiding in database.

To overcome misuse of data mining PPDM (Privacy Preserving Data Mining) was first introduced by Agrawal and Srikant in 1993.Aim of PPDM is to preserve privacy sensitive knowledge from disclosure. Many of the researchers have made their effort to preserve privacy for sensitive association rules.

The rest of the paper is organized as follows; in section 2 Association rules mining, Section 3, Association rule hiding, Section 4 Association rule hiding Approaches, Section 5 Literature Review, Section 6 Conclusion, Section 7 References.

Association rule mining

Association rule mining has been used in many application domains of data mining. Some applications are, finding patterns in medical database, business analysis, market analysis, extraction of information or knowledge from software metric, etc. It was first introduced by Agrawal et al in 1993[1].Let I = {I1, I2, …, Im} is a set of items, T = {T1, T2, … , Tn} is a set of transactions, Each of which contains items of the itemset I. Each transaction ti is a set of items such that ti ⊆ I. An association rule is an implication of the form: X →Y, where X ⊂ I, Y ⊂ I and X ∩Y = Ø.X (or Y) is a set of items, called itemset. In the rule X→Y, X is called the antecedent, Y is the consequent. It is obvious that the value of the antecedent implies the value of the consequent. The antecedent, also called the "left hand side" of a rule, can consist either of a single item or of a whole set of items. This applies for the consequent, also called the "right hand side", as well. Generally, only those item sets that fulfill a certain support requirement are taken into consideration. Support and confidence [2] are the two most important quality measures for evaluating the interestingness of a rule. The support of the rule X→Y is the percentage of transactions in T that contain X→Y. It determines how frequent the rule is applicable to the transaction set T. The support of a rule is represented by the formula,

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

$$\text{Support}(X \rightarrow Y) = \frac{|XUY|}{|D|} \quad (1)$$

Where |D| is total number of transactions in database D. Suppose the support of an item is 0.2%, it means only 0.2 percent of the transaction contain purchasing of this item.

The confidence of a rule is defined as percentage of the number of transaction that contains XUY to the total number of records that contains X. The confidence of the rule is represented by formula,

$$\text{Confidence}(X \rightarrow Y) = \frac{\text{Support } |XUY|}{\text{Support } |X|} \quad (2)$$

Where |X| denotes the number of transaction in database D that contains itemset X. Confidence is the measure of strength of association rule. Suppose the confidence of the association rule (X→Y) is 90%, it means that 90% of the transaction that contain X also contains Y together.

A rule X→Y is strong if support (X→Y) ≥ min_support and confidence (X→Y) ≥ min_confidence, where min_support and min_confidence [2] are two given minimum thresholds, min_support and min_confidence are two user defined values.

Association rule hiding techniques

- AIS Algorithm
- Apriori Algorithm
- FP-Tree Algorithm
- Dynamic Item Set Counting Algorithm
- Pincer Search Algorithm
- FApriori Algorithm

A.AIS Algorithm

AIS (Agrawal, Imielinski, and Swami) algorithm [1] was first algorithm for mining association rules in [Agrawal et al].There are two phases in the algorithm, first phase is to generate frequent item sets and second phase to generate confident and frequent association rule. This algorithm mainly developed to generate large item sets in a transaction database. In this algorithm only one item consequent association rules are generated, which means that the consequent of those rules only contain one item, for example we only generate rules like A∩B=C but not those rules as A=B∩C.

Main drawbacks of AIS, too many candidate item sets that finally turned out to be small are generated, which requires more space and wastes much effort that turned out to be useless [3].

B. Apriori Algorithm

Apriori algorithm [3] developed by Agrawal and Srikant is most widely used algorithm in Association rule mining. There are two phases in this algorithm; first phase is to generate the frequent item sets by generating the support count and minimum support of item sets. Second phase is to generate the rule by user defined parameter called minimum confidence. Apriori is the bottom up search algorithm, moving upward level wise in the lattice.

There are two drawbacks of apriori algorithm, one is the complex candidate set generation so most of the time space and memory is wasted in candidate set generation, second is multiple scan of database.

C. FP-Tree Algorithm

FP-tree [4] was developed by Han et al in 2000, FP-tree algorithm works like divide and conquer manner, there are two scan in the database, in the first scan list of frequent item sets generated by sorting frequency in descending order. In second scan the database is compressed into a FP-tree. Then FP-growth starts to mine the FP-tree for each item whose support is larger than minimum support by recursively building its conditional FP-tree. The algorithm performs mining recursively on FP-tree. The problem of finding frequent item sets is converted to searching and constructing trees recursively.

It is applicable to incremental database, because when new datasets arrives then repetition of whole process is done.

D. Dynamic Item Set Counting Algorithm

Dynamic Item Set Counting was proposed by bin et al in 1997, rationale of dynamic set counting algorithm is that it works like train pruning over data, with stops at intervals M between transaction file, it has made one pass over the data, and it starts all over again from the beginning for the next pass[5].

E. Pincer Search Algorithm

The pincer–search computing algorithm starts from the smallest set of frequent itemsets and moves upward till it reaches the largest frequent itemsets. The pincer-search algorithm is based on the principle of finding frequent item sets in a bottom up manner, but, at the

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

same time, it maintains a list of maximal frequent item sets [5].

In this algorithm, in addition to counting the supports of the candidate in the bottom-up direction, it also counts the support of the itemsets of some using in top up approach.

F.F Apriori Algorithm

Mihir et al [23], introduced FApriori algorithm, is a modified Apriori algorithm based on checkpoint. They proposed a method which can be combination of Apriori algorithm and reduced storage required to store candidate and execution time by reducing CPU time.

They introduced checkpoint concept based on support value to reduce execution time and storage space required to store candidate generated during scanning of datasets [23].

Association rule hiding appraoches

Data mining techniques has broader applications now days with some of data security issues. Association Rule hiding is used in PPDM to deal with data security issues. Association rule hiding is the process to convert the original database into sanitized database, so that sensitive association rule will be hidden.

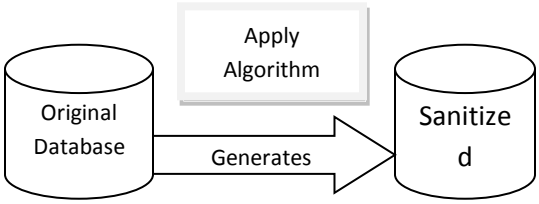

Fig.1 framework of Association Rule hiding

Many approaches have been proposed for Hiding Association Rules are classified into five categories:

1. Heuristic Based Approach

These approaches further divided into two groups:

1.1 Data Distortion Techniques

This technique permanently deletes some itemsets from the database, data distortion technique try to hide association rules by decreasing or increasing support to decrease or increase the support value is changed by new value. For example, it changes '1' to '0' or '0' to '1'. By decreasing or increasing support items in sensitive rule, or by changing value of confidence of items in sensitive rule.

M.Attallah et al. [7] proposed heuristic algorithms; they also give proof of NP-hardness of optimal sanitization [7].Verykios et al. [8] proposed five assumptions to hide sensitive knowledge in database. Y-H Wu et al. [9] proposed method to reduce the side effects in sanitized database, which are produced by other approaches [8]. K.Duraiswamy et al. [10] proposed a clustering based approach to reduce the time complexity of the hiding process.

1.2 Data Blocking Techniques

This technique replaces '1' and '0' with '?' in selected transaction of database, so entries in the database are not modified. So, it becomes very difficult for adversary to know the value of '?'.This technique is efficient than data distortion technique because it does not delete items permanently from database.

3. Border Based Approach

In this approach, lattice of the frequent and infrequent item sets is modified to hide sensitive association rules in original database. The itemsets between frequent and infrequent makes the border, border separate frequent and infrequent itemsets [6].Sun Yu [11] were the first to propose the border revision process. The authors in [12] proposed more efficient algorithms than proposed by Sun Yu [11].

4. Reconstruction Based Approach

These approaches are efficient than the Heuristic based approaches and generate less side effects than heuristic based approaches. In this approach first frequent item set is extracted from non frequent item set and privacy protected data is released. The new released data is then reconstructed from the sanitized knowledge base. This approach, first perform data perturbing and then reconstruct the database. Basically this approach reconstructs the database in a manner that all sensitive information has been hidden. This method cannot guarantee to find a consistent one within a polynomial time [6]. Y. Guo [13] proposed a FP tree based algorithm which reconstruct the original database by using non characteristic of database and efficiently generates number of secure databases.

4. Exact Based Approach

This approach was proposed by Gkoulalas and Verykios [14] for finding optimal solution for rule hiding problem. This approach rule hiding problem in to constraints satisfaction problem (CSP)and solve it by using binary integer programming (BIP).This approach is better but suffers from high time complexity to CSP.

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

5. Cryptographic Based Approach
Cryptographic based approaches mostly used in multiparty computation. The concept of secure multiparty computation was introduced in [15]. The idea of Secure Multiparty Computation is that a computation is secure if at the end of the computation, no party knows anything except its own input and the results. One way to view this is to imagine a trusted third party – everyone gives their input to the trusted party, who performs the computation and sends the results to the participants.

Vaidya and Clifton [16] proposed secure approach for sharing association rules when data are vertically partitioned. The authors in [17] proposed secure mining of association rules over horizontal partitioned data.

Literature review

A. Using Unknowns for hiding Predictive Association Rules
Shyue-Liang and Ayat Jafari [17] proposed two algorithms for hiding Sensitive Predictive Association Rules named ISL(Increase Support of L.H.S) and DSR (Decrease Support of R.H.S).First algorithm(ISL) tries to increase support of the left hand side of the rule. The Second algorithm (DSR) tries to Decrease Support of right hand side of the rule. So, the confidence of the association rule can be reduced. Either algorithm decrease or increase the support of the rule to hide he sensitive association rules.

B. Hiding Sensitive Association Rule by hiding Counters
The Ramesh Chandra Belwal et al [19] proposed a heuristic based algorithm for association rule hiding .they proposed a method for hiding sensitive association rule based on ISL (Increase the support of the item which is in the left hand side of the rule).they modified the definition of support and confidence, they introduced the use of a hidden counter in determining confidence and support New modified confidence and Support for the rule X →Y is,

$$\text{MConfidence} = \frac{|XUY|}{|X|+\text{Hidden Counter}} \quad [3]$$

$$\text{MSupport} = \frac{|XUY|}{|N|+\text{Hidden Counter}} \quad [4]$$

C. DSRRC (Decrease Support of R.H.S of Rule Cluster) Algorithm
Chirag Modi, Udai Pratap Roa, Dhiren Patel [17] proposed heuristic algorithm named DSRRC (Decrease Support of R.H.S of Rule Cluster).DSRRC hides the Sensitive Association rules by reducing support of R.H.S items of the rules. This algorithm only hides rules that contain single item on R.H.S of the rule. This algorithm does not maintain data quality and modifications on database was high.

D. MDSRRC (Modified Decrease Support of RHS of Rule Cluster)
Nikunj Domadiya, Udai Pratap Rao [20] introduced Heuristic Based algorithm MDSRRC (Modified Decrease Support of R.H.S of Rule Cluster).They proposed this algorithm to hide sensitive association rules with multiple items in consequent (R.H.S) and antecedent (L.H.S).This algorithm was proposed to overcome the limitations of DSRRC, it is able to hide multiple rules that in right hand side, with limited modification in database. In MDSRRC modifications on the database was less than DSRRC and it maintain data quality.

E. FHSAR (Fast Hiding Sensitive Association Rule) Algorithm
Chia-Chia Weng et al [21], introduced an algorithm, Fast Hiding Sensitive Association Rule (FHSAR), algorithm scans database single time which reduced the execution time. This algorithm is independent from size of database.

F. Hybrid Algorithm for Hiding Sensitive Association Rules
Niteen D. et al [22] introduced hybrid algorithm for hiding sensitive association rules. It uses combination of ISL and DSR and hides the association rules by modifying the database transactions so, confidence of the association rules can be modified. Memory Utilization of algorithm was less.

Conclusion and future work
In this paper we have studied various association rule mining techniques and approaches. We have discussed about various major algorithm for hiding sensitive association rules algorithms. Various algorithms are applicable to static database, when new dataset arrive these algorithms did not work efficiently. There is need to propose new algorithm for incremental datasets.

So, in future the focus will be on developing algorithm that will apply on incremental datasets. And memory utilization of such algorithm must be reduced in future.

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

References

1. R. Agarwal, T. Imielinski, and A. Swami, "Mining associations between sets of items in large databases". SIGMOD93, pages 207-216, Washington, D.C, USA, May 1993
2. Padam Gulwani "Association Rule Hiding by Positions Swapping of Support and Confidence". International journal of Information Technology and Computer Science, 2012.
3. Rakesh Agrawal and Ramakrishna Srikant, Fast Algorithms for Mining Association Rules in Large Databases, Proceedings of the Twentieth International Conference on Very Large Databases, pp. 487-499, Santiago, Chile, 1994.
4. Han, J. and Pei, J. 2000. Mining frequent patterns by pattern-growth: methodology and implications. ACM SIGKDD Explorations Newsletter.
5. Privacy preserving data mining based on association rule-a survey S.vijayrani,dr, Dr tamilrasi, R.seethalakshmi
6. Khyati B. Jadav, Jignesh Vania, Dhiren R. Patel "A Survey on Association Rule Hiding Methods" International Journal of Computer Applications, November 2013.
7. M. Atallah, A. Elmagarmid, M. Ibrahim, E. Bertino, and V. Verykios, "Disclosure limitation of sensitive rules," in Proceedings of the 1999 Workshop on Knowledge and Data Engineering Exchange, ser. KDEX '99. Washington, DC, USA: IEEE Computer Society, 1999, pp. 45–52.
8. V.S. Verykios, A.K. Elmagarmid, E. Bertino, Y. Saygin, and E. Dasseni, "Association rule hiding," IEEE Transactions on Knowledge and Data Engineering, vol. 16(4), pp. 434–447, April 2004.
9. Y. H. Wu, C.M. Chiang and A.L.P. Chen, "Hiding Sensitive Association Rules with Limited Side Effects," IEEE Transactions on Knowledge and Data Engineering, vol.19 (1), pp. 29–42, Jan. 2007.
10. K. Duraiswamy, and D. Manjula, "Advanced Approach in Sensitive Rule Hiding" Modern Applied Science, vol. 3(2), Feb. 2009.
11. X. Sun and P.S. Yu, "A Border-Based Approach for Hiding Sensitive Frequent Itemsets," In Proc. Fifth IEEE Int'l Conf. Data Mining (ICDM'05), pp. 426–433, Nov. 2005.
12. Moustakides and V.S. Verykios, "A Max-Min Approach for Hiding Frequent Itemsets," In Proc. Sixth IEEE Int'l Conf. Data Mining (ICDM'06), pp. 502–506, April 2006.
13. Y. Guo, "Reconstruction-Based Association Rule Hiding," In Proc. Of SIGMOD2007 Ph.D. Workshop on Innovative Database Research 2007(IDAR2007), June 2007.
14. A. Gkoulalas-Divanis and V.S. Verykios, "An Integer Programming Approach for Frequent Itemset Hiding," In Proc. ACM ConfInformation and Knowledge Management (CIKM '06), Nov. 2006.
15. Yao, A. C.-chih. (1986). "How to Generate and Exchange Secrets". Exchange Organizational Behavior Teaching Journal, (1), 162-167
16. J. Vaidya and C. Clifton, "Privacy preserving association rule mining in vertically partitioned data," In proc. Int'l Conf. Knowledge Discovery and Data Mining, pp. 639–644, July 2002.
17. Shyue-Liang Wang, Ayat Jafari "Using Unknowns for Hiding Sensitive Predictive Association Rules" IEEE 2005
18. Chirag Modi, Udai Pratap Roa, Dhiren Patel "Maintain Privacy and Data Quality in Privacy Preserving Association Rule Mining" IEEE 2010
19. Ramesh Chandra Belwal, Jitendra Varshney, Sohel Ahmed Khan, Anand Sharma, Mahua Bhattacharya, "Hiding Sensitive Association Rules Efficiently By Introducing New Variable Hiding counter", IEEE International conference on Service Operations, Logistics and informatics, Vol.1,Oct. 2008, pp 130-134
20. N Domadiya and U. P. Rao, "Hiding Sensitive Association Rules to Maintain Privacy and Data Quality in Database" 2013 3rd IEEE International Advance Computing Conference (IACC), pp. 1306-1310, 2013.
21. Chih-Chia Weng, Shan-Tai Chen & Hung-Che Lo "A Novel Algorithm for Completely Hiding Sensitive Association Rules" IEEE Intelligent Systems Design and Applications, 2008.
22. Niteen Dhutraj, Siddhart Sasane, Vivek Kshirsagar "Hiding Sensitive Association Rule For Privacy Preservation", IEEE Transactions on Knowledge and Data Engineering Year 2013.
23. Mihir R.Patel, Dipti P.Rana, Rupa G.Mehta, "FApriori: A modified Apriori Algorithm Based on checkpoint" 2013 International Conference on Information System and Computer Network, IEEE 2013