A Survey of Various Methodologies for Hiding Sensitive Association Rules

Tapan Sirole
Department of computer Science
MANIT, Bhopal (MP), India

Jaytrilok Choudhary
Department of Computer Science
MANIT, Bhopal (MP), India

ABSTRACT
Data mining and hiding are the future research directions in the field of knowledge engineering. The main challenges in data mining are finding the sensitive association and hide them without revealing sensitive information. The association rule hiding is a process in which the original database is modified in such way that precise sensitive rules are disappeared. In this paper, a survey of various recent approaches of association rule hiding has been described along with the comparison between them.

General Terms
Privacy Preserving Data mining (PPDM). Association rule, Sensitive Association rules.

Keywords
Association rule hiding, Association rule, data mining, Privacy preservation.

1. INTRODUCTION
Today is the generation of data, there is huge amount of data being produced every day from different resources. It is estimated that the amount of data stored in different database is almost doubled in every two years, this data store in storage devices in form of raw data. Thus, there is a need of some techniques in order to extract useful pattern or information from stored data. The Data mining is the technique that extracts the knowledge from the large volume of data. Basically data mining is a process for analysing data from different prospective and generating some useful information, the extracted information may be used to grow up the business by different organization, for example by extracting the knowledge form market basket database. The market owners may increase their revenue by offering many exciting offer for customer, the extracted information from the data may contain sensitive information like purchasing habits of customer, confidential data of some organization etc.

Besides extracting information or knowledge from raw data, there is also need for some technique or scheme that deal with security of that information, privacy preserving in data mining (PPDM) is the technique that deal with the security of the information that extracted by data mining techniques, PPDM allow to mine the information from large amount of data while protecting sensitive information defined by the data base owner, or the information that database owner do not want to disclose. The main aim of PPDM is to minimized the risk of misuse of data while does not affect the data mining techniques. Privacy preserving data mining is first introduced by Agrawal and Srikant[1].

The rest of the paper is organized as follows. In second section discussion about the Association rule mining. The association rule hiding has been discussed in section 3. Then, the various strategies of association rule hiding. The literature review of association rule hiding has been discussed in section 5. Finally, concluded the survey on sensitive association rules hiding.

2. ASSOCIATION RULE MINING
Association rule mining is important field to be considered under privacy preserving data mining, Agrawal et al. [2] was first who proposed the concept of association rule mining, basically association rule is the if-then relationship among the data. Consider example for better understanding the concept of association rule "If a customer buys a dozen eggs, then he is 85% likely to also purchase milk". By analysing above example it can be concluded that milk is somewhat related to egg because every time a customer buy egg, 85% of the time he/she also likely to buy milk. Initially association rule is used for market basket analysis in order to determine the relationships among the items bought by customers. Association rule is of the form X→Y, where X any Y is belongs to collection of items set and the intersection of X and Y must be null. Every association rule must satisfy two constraint support and confidence.

Support of a rule X→Y is the percentage of transactions of the transaction database that contain item XUY. Support for the rule (X→Y) can be calculated by using the formula given in (1)

$$support(X \rightarrow Y) = \frac{|XUY|}{N} \quad (1)$$

Where N is total number of transaction in transactional database.

Confidence of a rule X→Y is the percentage of transactions in the transaction database that contain X also contain Y. the confidence of rule (X→Y) can be calculated by using following formula

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

3. ASSOCIATION RULE HIDING
The association rule hiding is one of the techniques that used in PPDM. The association rule hiding methodologies aim at sanitizing the original database in a way that at least one of the following goals is accomplished [3].

1. No rule that 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 be also revealed from the sanitized database, when this database is mined at the same or at higher thresholds.
2. All the non sensitive rules that appear when mining the original database at pre-specified thresholds of confidence and support can be successfully mined from the sanitized database at the same thresholds or higher.

3. No rule that was not derived from the original database when the database was mined at pre-specified thresholds of confidence and support can be derived from its sanitized counterpart when it is mined at the same or at higher threshold.

Association rule hiding process totally depend on support or confidence of the rule, there is two way to hide any rule ,either decrease support up to certain threshold or decrease confidence up certain threshold, so the mining algorithm, that works on support not able to mine sensitive rules. if we analyse the basic equation for finding support and confidence given in eq.(1) and (2) we can find that we have two option to decrease the support and confidence of any rule

1. By decreasing the numerator item support while keeping the support of denominator unchanged.

2. By increasing support of denominator items while keeping the support of numerator items unchanged.

Fig.1 is showing general framework for association rule hiding. However the modification on the database may cause some side effect that may lead to some disturbance in association rule mining, following are the some side effect that may occurs in the process of rule hiding:

Lost Rules: the non sensitive association rules which are present in original database and can be mined by applying mining algorithm but cannot be mined after applying hiding algorithm from modified database.

False rules: the sensitive association rules which are not hidden by hiding algorithm and can be mine by applying mining algorithm on modified database.

Ghost rules: the rules which are not present in original database but generated after applying hiding algorithm.

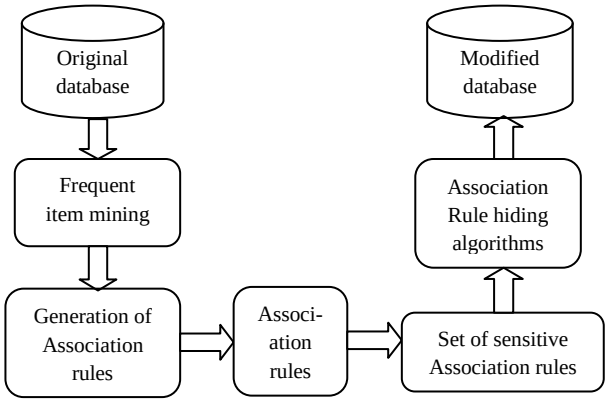

Fig.1 General Framework for hiding sensitive association rule

4. ASSOCIATION RULE HIDING STRATEGIES

The strategies of Association rule hiding are classified as follows

4.1 Heuristic based approach

This approach hides sensitive association rules by using two methods

4.1.1 *Data Distortion based technique*

M. Attallah et. al[3] were first use this technique for hiding association rules, they also gave the proof of NP hardness of optimal sanitization. In this technique rules are hiding by modifying database matrix by changing the value of some items in database matrix by 0 to 1 or vice versa. The data distortion technique contains two basic methods for hiding rules. In first method, rules are hide by decreasing the support of the rule up to an acceptable level and in second method, the confidence of the rule is reduced up to certain threshold. Table 1 Showing basic data distortion technique.

Verykios et al. [4] proposed five different algorithms for hiding association rules. Three of them based on reduce support and remaining two are based on reducing confidence up to an acceptable level.

Table 1. Hiding A→D by Distortion

A	B	C	D
1	1	0	**1**
1	0	0	1
1	0	1	1
0	1	1	0
1	1	1	1
0	0	1	0

→

A	B	C	D
1	1	0	0
1	0	0	1
1	0	1	1
1	1	1	0
1	1	1	1
0	0	1	0

4.1.2 *Data blocking based techniques*

Y. Saygin et al.[5] and [6]have proposed algorithm for hiding sensitive association rules based on data blocking technique. In this technique, rules are hide by changing the value of some item in database matrix from 0 or 1 to ?(unknown).So, the support of certain items goes down to certain level and rule mining algorithm not able to mine the sensitive rules. Table 2 Showing basic data blocking technique.

Table 2. Hiding A→D by Blocking

A	B	C	D
1	1	0	1
1	0	0	**1**
1	0	1	1
0	1	1	0
1	1	1	1
0	0	1	0

→

A	B	C	D
1	1	0	1
1	0	0	?
1	0	1	1
0	1	1	0
?	1	1	1
0	0	1	0

4.2 Border based Approach

The border based approach hide sensitive association rules by modifying the border in the lattice of frequent and infrequent item sets of the original database. The item set between frequent and infrequent items make the border. The border consist the item sets which separate the frequent item set from infrequent item set. Sun and Yu [7] were first who introduce the concept of border.

4.3 Reconstruction based Approach

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.

4.4 Exact Approach

This is a non heuristic algorithm which formulates the rule hiding problem in to constraint satisfaction problem or optimal problem which is solved by integer programming. Divanis and Verykios[8] were first who used the exact approach for hiding rules. It provides an optimal solution for the problem of association rule hiding.

4.5 Cryptographic approach

This technique used in multi-party computation where data is distributed in different location. The database owner may want to share their data, but at the same time they want to ensure their privacy at their end. Cryptographic approach can be categorized in two categories vertical partitioned distributed data and horizontal partitioned distributed data.

In horizontal partitioning different rows are placed in different tables that are distributed in different locations. In vertical partitioning some columns kept in one table and remaining column kept in other tables.

A comparatively analysis of different hiding approaches given in table 3.

Table 3. Analysis of different association rule hiding strategies.

APPROACH		ADVANTA-GES	LIMITATIO-N
Heuristic approach	(Data Distortion)	More efficient, scalable	Difficult to revert the changes made in database
	(Data Blocking)	It maintains veracity of database, since instead of inserting false value it just block original value.	Suffer from various side effects like ghost rule, lost rule etc.
Border based approach		Maintain database quality by selecting the transaction that produces minimal side effect.	Theory of border difficult to understand Based on heuristic approach.
Exact approach		Provides an optimal solution without any side effects.	High complexity due to linear integer programming.
Reconstruction based approach		Lesser side effect than heuristic based approaches.	Number of transaction is restricted in new released database.
Cryptography based approach		Provide security in multi party computation or where data distributed in different locations.	Does not provide security for the output of the computation.

5. LITERATURE REVIEW

5.1 ISL(Increase support of LHS) and DSR(Decrease Support of RHS)

Shyue-Liang Wang and Ayat Jafari[10] proposed two algorithm for hiding sensitive predictive association rules, namely ISL and DSR, the ISL algorithm hide the rule by trying to increase the support of left hand side items of the rule. The DSR algorithm hide rule by decreasing the support of right hand side items of the rule. Both algorithms hide the rules by either decreasing or increasing the support of those items that participated in the sensitive rule.

5.2 Association rule hiding using hidden counter

The Ramesh Chandra Belwal et al.[11] proposed a heuristic based algorithm for addressing the problem of 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 support and confidence New modified confidence for the rule X → Y is

$$MConfidence = \frac{|XUY|}{|X|+HIdden\ Counter} \quad (3)$$

New modified support for the rule X → Y is

$$Msupport = \frac{|XUY|}{N+HIdden\ Counter} \quad (4)$$

5.3 DSSR (Decrease support of R.H.S items in rule clusters)

C. N. Modi, U. P. Rao et al. [12] proposed a algorithm for hiding sensitive association rules based on heuristic approach, proposed algorithm named DSSR(Decrease support of R.H.S items in rule clusters) the algorithm is based on DSR (Decrease support of right hand item) ,the algorithm hide association rule my making the cluster of sensitive association rule based on right hand items, then calculates the sensitivity of each cluster ,the sensitivity of cluster is the sum of sensitivity of each item present in the cluster, then index sensitive transactions for each cluster and sorts all the clusters decreasing order of their Sensitivities, then the hiding process hide the rule by deleting common R.H.S. item of the rules in cluster, from the sensitive transactions.

The advantage of proposed algorithm is that it maintain database quality do not make major changes in the database, the disadvantage of proposed technique is it can hide only those sensitive association rules which contain only single item in the right hand side of the rule.

5.4 MDSRRC (Modified Decrease Support of R.H.S item of Rule Clusters)

N Domadiya et al.[13] proposed a heuristic based algorithm for hiding sensitive association rule, the algorithm named MDSRRC (Modified Decrease Support of R.H.S. item of Rule Clusters) basically the proposed algorithm is the modification of algorithm DSRRC proposed in [8],the proposed algorithm overcome the limitation of DSRRC, it is able to hide the sensitive association rule that contain multiple items in right hand side.

The main advantage of proposed algorithm is, it does not make major changes in the database and it also able to hide rule which contain multiples item in right hand side of the rule

5.5 FHSAR (Fast Hiding Sensitive Association Rules)

Chih-Chia Weng et al. [14] proposed an algorithm for hiding sensitive association rules, the proposed algorithm named FHSAR (Fast Hiding Sensitive Association Rules) can hide all sensitive association rules without modifying the given database, the proposed algorithm also very efficient in terms of time because it scan database only once in order to hide sensitive rules, the proposed algorithm also independent of size of database

5.6 ISLRC (Increase Support of L.H.S. item of Rule Clusters)

Sanjeev Keer et al. [15] proposed a heuristic based approach for hiding association rule, the proposed algorithm hide the association rules that contain multiple items on right hand side , proposed method hide sensitive association rule by creating clusters of rules based on common LHS items of the rule. Then the sensitive transactions for all cluster is indexed, then the sensitivity of each cluster is calculated, the sensitivity of cluster is the sum of sensitivities of each items, then clusters are sorted ,and finally rule hiding process hide the rules.

The main advantage of proposed method is it minimized the side effect on the database and reduced time complexity by creating the clusters, limitation of the proposed method is it cannot hide the rules that contain multiple items on LHS.

6. CONCLUSION

Association rule hiding is a technique for hiding sensitive information in database. It is one of the techniques used in PPDM. In this paper, the various techniques of association rule hiding have been discussed. The comparative study, including advantages and limitations, of each technique also has been reviewed. In future, some new techniques can be found by combining different hiding techniques to reduce the side effect on database and time complexity for large amount of data.

7. REFERENCES

[1] Agarwal and Srikant ,"Privacy-preserving data mining", *In ACM SIGMOD*, May 2000, pp. 439-450 .

[2] R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules between Sets of Items in Large Database," Proceedings of the ACM SIGMOD Conference on Management of Data, Washington, D.C., USA, pp. 207-216, 1993

[3] M. Atallah, E. Bertino, A. Elmagarmid, M. Ibrahim, and V. S.Verykios "Disclosure limitation of sensitive rules."In Proc. of the 1999 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX'99), pp. 45–52, 1999.

[4] V.S. Verykios, A. Elmagarmid, E. Bertino, Y. Saygin, and E. Dasseni, "Association rule hiding," IEEE Transactions on Knowledge and Data Engineering, Vol. 16, No.4, 434–447, 2004 .

[5] Y. Saygin, V. Verykios, and C. Clifton, "Using Unknowns to Prevent Discovery of Association Rules" ACM SIGMOD, Vol. 30, No. 4, pp. 45–54, 2001.

[6] Y. Saygin, V. Verykios, and A. Elmagarmid, "Privacy preserving association rule mining," In: Proc. Int'l. Workshop on Research Issues in Data Engineering (RIDE 2002), pp.151–163, 2002.

[7] X. Sun, and P. Yu, "A Border-Based Approach for Hiding Sensitive Frequent Itemsets," In: Proc. Fifth IEEE Int'l. Conf. Data Mining (ICDM 2005), pp. 426–433, 2005.

[8] A. Gkoulalas-Divanis, and V. S. Verykios, " Exact knowledge hiding through database extension" IEEE Trans Knowledge Data Eng 2009, pp. 699–713.

[9] Gkoulalas-Divanis, Aris, Verykios, Vassilios S. Association Rule Hiding for Data Mining, Springer Series: Advances in Database Systems, Vol. 41, 1st Edition., 2010, p.13.

[10] Shyue-Liang Wang and Ayat Jafari Using Unknowns for Hiding Sensitive Predictive Association Rules IEEE 2005.

[11] 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.

[12] C. N. Modi, U. P. Rao, and D. R. Patel, "Maintaining privacy and data quality in privacy preserving association rule mining," 2010 Second Iternational conference on Computing,Communication and Networking Technologies, pp. 1–6, Jul. 2010.

[13] 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.

[14] 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.,vol 3, pp.202-208, 2008.

[15] Sanjeev Keer and prof. Anju Singh, Hiding Sensitive association Rule using cluster s of sensitive association rule, International Journal of computer science and Network vol.1 Issue 3,June 2012.