Semiconductor Optoelectronics
ISSN: 1001-5868

DEVELOPMENT OF NOVEL METHOD TO HIDE SENSITIVE ASSOCIATION RULES

Ashoktaru Pal and Dr. Lokendra Singh Songare

Department of Computer Science & Engineering, Dr. A.P.J. Abdul Kalam University, Indore (M.P.), India

Abstract:

The information remains safe and unthreatened against unwarranted series of actions is the most important movement now-a-days. Data mining or knowledge discovery in databases has been developed as an important technology for identifying knowledge from large quantities of data. It is possible to infer sensitive information, including personal information, or even patterns from non-sensitive information. There is a need to prevent disclosure not only of confidential personal information from summarized or aggregated data, but also prevent data mining techniques from discovering sensitive knowledge which is not even known to the database owners. Association Rule Hiding in Privacy Preserving Data Mining can be achieved through formative database called sanitized database from the original database in centralized database environment. In this article, development of novel method to hide sensitive association rules has been discussed.

Keywords: Sensitive, Association, Novel, Development

INTRODUCTION:

Data sanitization is the process of hiding sensitive information in test and development databases. It is accomplished by modifying it with original view but unreal data of a similar type. The disclosed data must be sanitized in order to protect valuable business information. There is a legal obligation in most countries, to do the sanitization process. [1] The Association Rules are derived from the original dataset. It is categorized as weak and strong association rules. The confidence of the weak association rules is lower than the user specified threshold and the strong rules are higher. The strong association rules are categorized as sensitive and non-sensitive association rules. [2] It is shown in Figure 1. The sensitivity is decided by the data owner.

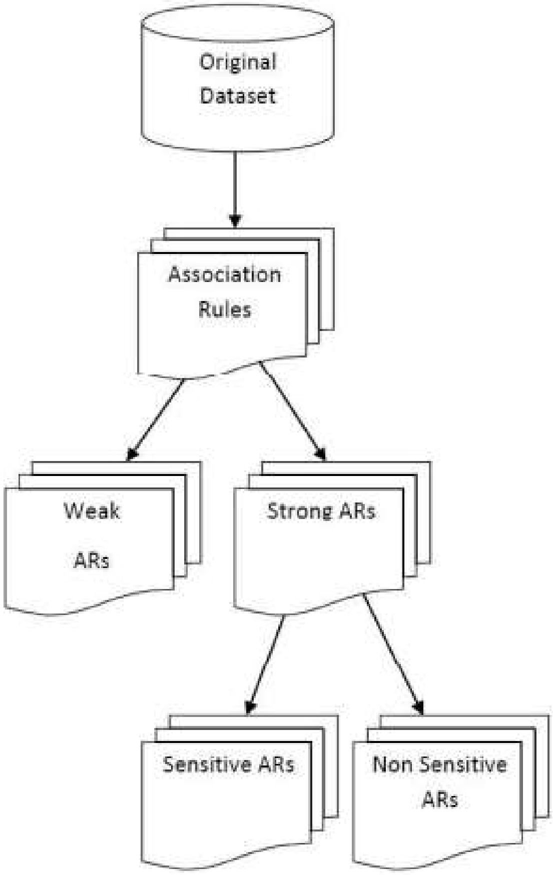

Figure 1. Association Rules

The Association Rule Hiding (ARH) algorithm is implemented over the original dataset to get the sanitized dataset. The side effects such as lost rules, ghost rules, transaction modification and hiding failures occur due to modification as shown in Figure 2. The strong association rules of the original dataset are compared with sanitized dataset for analyzing the side effects. Most of the existing association rules hiding techniques are enduring from the side effects. [3] The above-mentioned side effects play an important role in the motivation of proposed method.

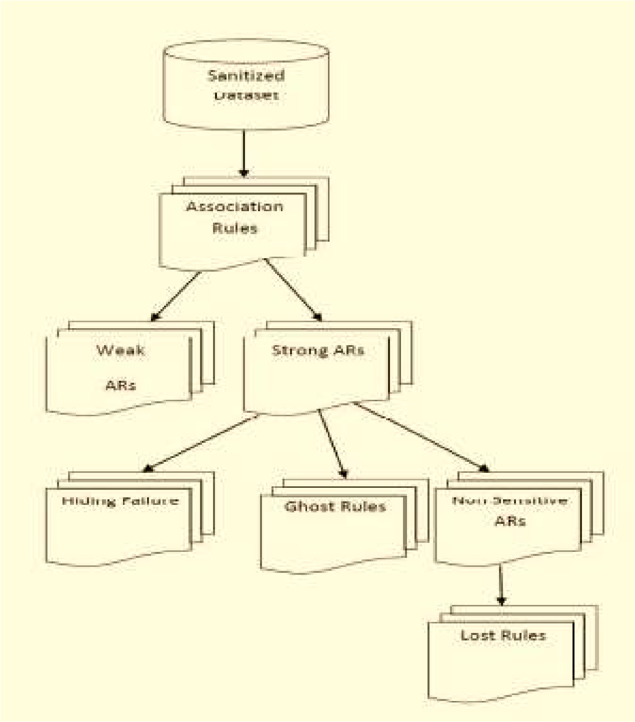

Figure 2. Side Effects of ARH

PROPOSED METHOD:

All the rules that contain a particular item or selected rules may be considered as sensitive by the data analyst. Some of the existing association rule hiding algorithms hide all the rules which contain the input items. The others hide the selected sensitive rules. The proposed method has been given either items or rules as input. If the items are given, all the rules having the specified items are hidden. Some of the selected rules are also hidden by using this proposed method. The transactions in the input dataset are modified to hide the sensitive rules. [4] Thus the output of the proposed method is the modified dataset. All the association rules are extracted using Association Rule Mining algorithm. Exemplary Association Rule (EAR) is formed using the sensitive rules and all the rules. The Left Hand Side (LHS) and the Right Hand Side (RHS) of EAR is used to prefer the set of transactions in the original dataset for modification.

The proposed method adopts two approaches:

- Heuristic Sensitive Association Rule Hiding (HSARH)
- Genetic Based Sensitive Association Rule Hiding (GBSARH)

HSARH hides all the rules contains the input item and GBSARH hides the selected rules. The steps involved in the proposed method are as follows:

Step 1: Original Dataset, Minimum Support, Minimum Confidence and Set of Items or Set of Rules are given as input.

Step 2: For each item in the set, Exemplary Association Rule is formed and HSARH is employed over the Original Dataset to obtain the modified dataset.

Step 3: For each rule in the set, the LHS and RHS are combined to form Exemplary Association Rule. GBSARH is applied over the Original Dataset to get the modified dataset.

HSARH hides more than one rule at a time and GBSARH hides one rule at a time.

EXEMPLARY ASSOCIATION RULE (EAR):

EARs are least set of rules satisfying specified constraints. EAR is generated for the input items as well as rules.

- Items Association Rules are generated using Association Rule Mining algorithm. All the association rules containing the input item are selected. Exemplary association rules are formed by having the input items on its left-hand side; all its consequents of the selected rules are combined on the right hand side. For every input item, EAR is generated.

- Rules The LHS and RHS of the input rule are combined and it is taken as the LHS of EAR. ARM generates the association rules. The rules containing the LHS of EAR are selected. The consequent items of the selected rules are considered as the RHS of EAR. For every input rule, EAR is generated.

HEURISTIC SENSITIVE ASSOCIATION RULE HIDING (HSARH):

The research work is initiated with heuristic approach to hide all the sensitive rules of specified items. Though the existing algorithms using heuristic approaches are very effective and fast, there are several circumstances in which these algorithms suffer from undesirable side effects that lead them to identify approximate hiding solutions. A new heuristic approach is proposed to improve the quality of identified hiding solution. [5]

The proposed heuristic approach selects a set of transactions for modification through Exemplary Association Rules (EAR).

The transactions containing the input items in LHS and RHS of EAR fully or partially are considered as EAR supporting transactions. The transactions which do not contain the items in LHS and RHS of EAR are considered as EAR non-supporting transactions. EAR supporting transactions are selected for modification.

The sanitization technique is applied in order to maintain the quality and privacy of the dataset. The proposed approach sanitizes the dataset in two ways:

i. Swapping In this operation, the input item is removed from EAR supporting transaction and it is added to the EAR no supporting transaction. The advantage of swapping is that the occurrence of the items remains same in both the original and sanitized dataset.

ii. Deletion The input item is removed from EAR supporting transaction. Minimal side effect occurs while deleting the items.

The selection of sanitization technique is based on the number of EAR supporting and non-supporting transactions. If the number of Exemplary association rules supporting transactions is lesser than non-supporting transactions, then swapping is possible. If not, the items are deleted from the supporting transaction. [6] This approach can hide multiple rules per iteration. The sequence of steps involved in heuristic approach is shown in Figure 3.

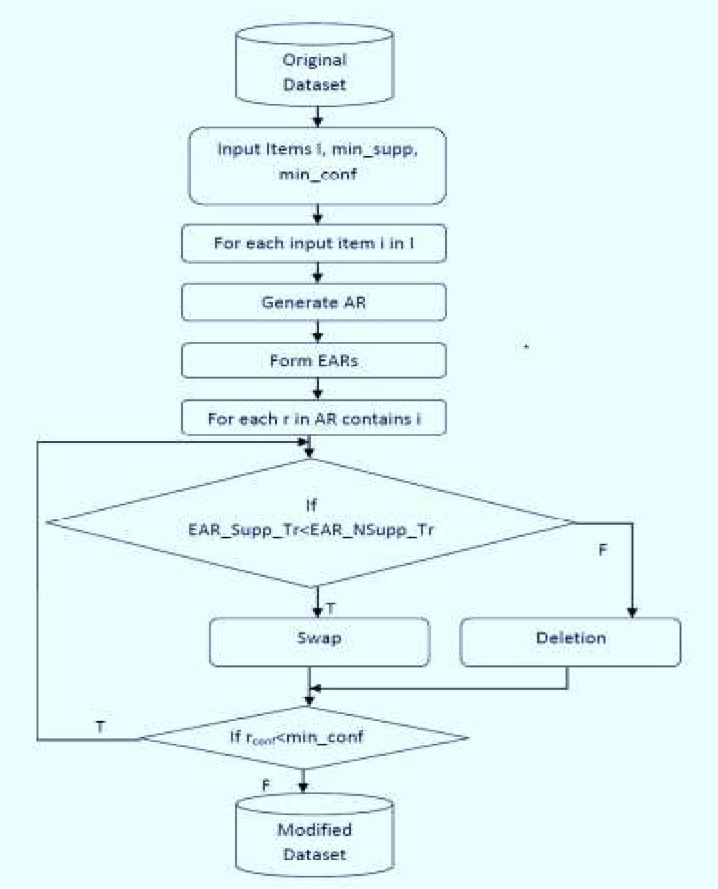
Original Dataset
Input Items I, min_supp, min_conf
For each input item i in I
Generate AR
Form EARs
For each r in AR contains i
If EAR_Supp_Tr<EAR_NSupp_Tr
F
T
Swap
Deletion
T
If r_{conf}<min_conf
F
Modified Dataset

Figure 3. Heuristic Approach

GENETIC BASED SENSITIVE ASSOCIATION RULE HIDING (GBSARH):

GBSARH hides the selected sensitive rules with a Genetic Algorithm (GA). EAR is generated using the items in selected sensitive rule. The transactional dataset D is assumed as Initial Population. D is transformed into Binary Dataset. The presence of item is represented as ‘0’ and absence as ‘1’.

The fitness value for every transaction in D is computed by using equation (3.1). The optimality of solution depends on the complexity of fitness function. The possible strength of fitness function ensures a desirable level of optimal solution.

$$\textbf{Fitness Value } fv=\sum_{i=1}^{j} S_i+\left|\frac{1}{\sum_{k=1}^{l} NS_k}\right|-\sum_{m=1}^{n} ERHS_m \qquad \text{------- (4.1)}$$

Si..j-Sensitive Items in the transaction.

NSk..l - Non Sensitive Items in the transaction (except ERHS).

ERHS – Exemplary Association Rule’s Right Hand Side.

The fitness function is proposed in such a way which reduces the side effects of association rule hiding technique. The fitness value is determined by the number of sensitive items and non-sensitive items and RHS of EAR occurring in the transaction.

The proposed fitness function has been designed based on the following three components:

- The first component shows that the maximum number of items in transaction. It reduces the hiding failure.
- The second component represents the minimum number of non-sensitive items. It minimizes the ghost rule side effect.
- The last one shows that there is minimum number of RHS of EAR in transaction. It reduces the lost rule side effects.

The transaction having higher fitness value is to be selected for modification.

The following Genetic operations are applied over the dataset:

TOURNAMENT SELECTION:

In Tournament Selection, the two transactions are randomly selected from the dataset. The more fit of these two transactions are selected for mating pool.

SINGLE POINT CROSS-OVER:

In single point crossover, the transaction is split in to two portions such as head and tail. The head of one transaction is combined with the tail of another transaction in the mating pool as illustrated in Figure 4

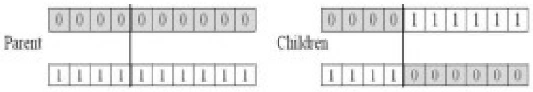

Figure 4. Single Point Cross Over

RANDOM MUTATION:

This operation randomly changes from 0 to 1 or 1 to 0. This allows the transactions in the dataset from becoming similar to each other.

The fitness value of offspring transactions are also computed. The transaction containing maximum fitness value in input dataset is replaced with offspring transaction containing minimum value. The process is repeated until the side effects in the modified dataset are admitted in least amount. The steps in the Genetic approach is shown in Figure 5.

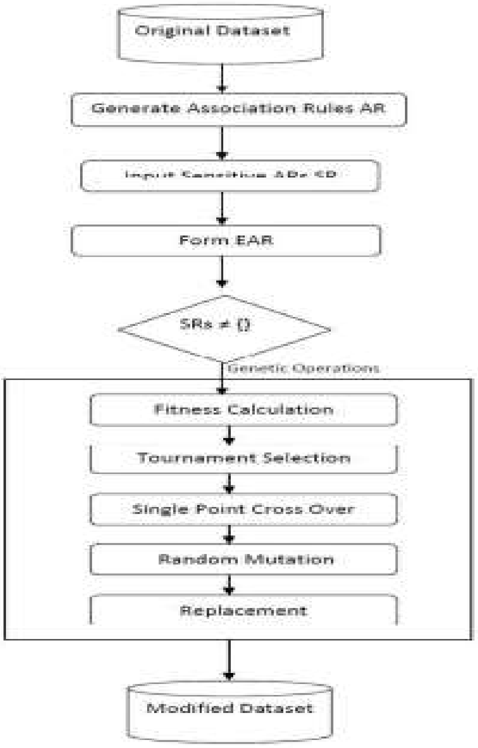

Figure 5. Genetic Approach

CONCLUSION:

The proposed method for hiding sensitive association rule is presented. The proposed method adopts HSARH and GBSARH approaches. The method can be given item or rule as input and a source dataset with minimum support and confidence. Modified dataset is the output of the proposed method. [7] Association Rule mining technique is applied over the modified dataset. The rules that contain the input item or the rules are not disclosed. Hence, privacy is preserved while mining the dataset. [8] All the sensitive rules are hidden in sample dataset.

REFERENCES:

1. Islam, M. Z. and Brankovic, L. (2011), ‘Privacy preserving data mining: A noise addition framework using a novel clustering technique’, Knowledge-Based Systems 24(8), 1214–1223.

2. Jalla, Hanumantha & Girija, P. (2016). A Novel Approach for Horizontal Privacy Preserving Data Mining. 10.1007/978-81-322-2752-6_9.

3. Giannotti, F., Lakshmanan, L. V., Monreale, A., Pedreschi, D. and Wang, H. (2012), 'Privacy-preserving mining of association rules from outsourced transaction databases', IEEE Systems Journal 7(3), 385–395.

4. Bhaladhare, P.R. and Jinwala, D.C., 2016. Novel Approaches for Privacy Preserving Data Mining in K-anonymity Model. J. Inf. Sci. Eng., 32(1), pp.63- 78. 25

5. Eswaran, Poovammal & Ponnavaikko, Murugesan. (2009). An Improved Method for Privacy Preserving Data Mining. 1453 - 1458. 10.1109/IADCC.2009.4809231.

6. Chen, K, & Liu, L 2011, 'Geometric data perturbation for privacy preserving outsourced data mining', Knowledge and information systems, vol. 29, no. 3, pp. 657-695.

7. Alpa Shah and Ravi Gulati. Article: Privacy Preserving Data Mining: Techniques, Classification and Implications - A Survey. International Journal of Computer Applications 137(12):40-46, March 2016.

8. Fan C, Xiao F, Li Z, Wang J. Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: A review. Energy and Buildings. 2018 Jan 15;159: pp. 296-308.