

# A Well-Organized Method for Protecting Sensitive Information

Priyanka Choubey<sup>1</sup>, Savita Rathoad<sup>2</sup>

<sup>1</sup>M. Tech. (Final Year) Truba College of Engineering & Technology, Indore, India

<sup>2</sup>Professor, Truba College of Engineering & Technology, Indore, India

**Abstract:** *Data mining services need correct input data for their results to be significant, but privacy concerns may power users to offer spurious information. To protect client privacy in the data mining process, a multiplicity of techniques based on random perturbation of data records have been proposed lately. One known fact which is very significant in data mining is discovering the association rules from database of transactions each transaction consists of set of objects. Two important terms support and confidence are linked with each of the association rule. Actually any rule is called as sensitive if its disclosure risk is above a certain privacy threshold. Sometimes we do not want to disclose sensitive rules to the public because of privacy purposes. There are many approaches to hide confident association rules which take the support and confidence as a base for algorithms and many more. The proposed work has the basis of reduction of support and confidence of sensitive rules but this work is not restriction or disturbing the given database of transactions directly. The proposed algorithm uses some customized definition of support and confidence so that it would hide any desired sensitive association regulation without any side effect. In reality the enhanced technique is by means of the same method (as previously used method) of receiving association rules but modified definitions of support and confidence are used*

**Keywords:** Association rules Support, Confidence, Data mining, Data hiding, etc.

## 1. Introduction

Many government agencies, businesses and non-profit organizations in order to support their short and long term planning activities, they are searching for a way to collect, store, analyze and report data about individuals, households or businesses. Information systems, therefore, contain confidential information such as social security numbers, income, credit ratings, type of disease, customer purchases, etc., that must be properly protected. Let us suppose that we are negotiating a deal with Dedtrees Paper Company, as purchasing directors of Big Mart, a large supermarket chain.

They offer their products in reduced price, if we agree to give them access to our database of customer purchases. We accept the deal and Dedtrees starts mining our data. By using an association rule mining tool, they find that people who purchase skim milk also purchase Green paper.

Dedtrees now runs a coupon marketing campaign saying that “you can get 50 cents off skim milk with every purchase of a Dedtrees product”. This campaign cuts heavily into the sales of Green paper, which increases the prices to us, based on the lower sales. During our next negotiation with Dedtrees, we find out that with reduced competition they are unwilling to offer us a low price. Finally, we start to lose business to our competitors, who were able to negotiate a better deal with Green paper. The scenario that has just been presented, indicates the need to prevent disclosure not only of confidential personal information from summarized or aggregated data, but also to prevent data mining techniques from discovering sensitive knowledge which is not even known to the database owners. The necessity to combine the confidentiality and the legitimate needs of data users is imperative. Every disclosure limitation method has an impact, which is not always a positive one, on true data values and relationships. Ideally, these effects can be quantified so that their anticipated impact on the

completeness and validity of the data can guide the selection and use of the disclosure limitation method. The hiding strategies that we propose are based on reducing the support and confidence of rules that specify how significant they are. In order to achieve this, transactions are modified by removing some items, or inserting new items depending on the hiding strategy. The constraint on the algorithms is that the changes in the database introduced by the hiding process should be limited, in such a way that the information loss incurred by the process is minimal. Selection of the items in a rule to be hidden and the selection of the transactions that will be modified is a crucial factor for achieving the minimal information loss constraint. We also perform a detailed performance evaluation and validation study in order to prove that the proposed algorithms are computationally efficient, and provide certain provisions on the changes that they impose in the original database. According to this, we try to apply minimal changes in the database at every step of the hiding algorithms that we propose.

## 2. Background and Related Work

Data mining is the process of extracting hidden patterns from data. As more data is gathered, with the amount of data doubling every three years, data mining is becoming an increasingly important tool to transform this data into knowledge. It is commonly used in a wide range of applications, such as marketing, fraud detection and scientific discovery. Data mining can be applied to data sets of any size, and while it can be used to uncover hidden patterns, it cannot uncover patterns which are not already present in the data set.

Data mining extracts novel and useful knowledge from data and has become an effective analysis and decision means in corporation. Two problems are addressed in PPDM: one is the protection of private data; another is the protection of sensitive rules (knowledge) contained in the data. The former

settles how to get normal mining results when private data cannot be accessed accurately; the latter settles how to protect sensitive rules contained in the data from being discovered, while non-sensitive rules can still be mined normally. The latter problem is called knowledge hiding in database in (KHD) which is opposite to knowledge discovery in database (KDD). And association rule hiding problem we focus is one of problems in KHD. In data mining, association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. Piatetsky-Shapiro describes analyzing and presenting strong rules discovered in databases using different measures of interestingness. Based on the concept of strong rules, Agrawal et al introduced association rules for discovering regularities between products in large scale transaction data recorded by point-of-sale (POS) systems in supermarkets. For example, the rule Found in the sales data of a supermarket would indicate that if a customer buys onions and potatoes together, he or she is likely to also buy beef. Such information can be used as the basis for decisions about marketing activities such as, e.g., promotional pricing or product placements. In addition to the above example from market basket analysis association rules are employed today in many application areas including Web usage mining, intrusion detection and bioinformatics "Let  $D$  be a database,  $R$  be the set of rules mined from  $D$  based on a minimum support threshold  $\_$ , and  $RR$  be a set of restrictive rules that must be protected according to some security/privacy policies. The goal is to transform  $R$  into  $R\_$ , where  $R\_$  represents the set of non-restrictive rules. In this case,  $R\_$  becomes the released set of rules that is made available for sharing. Ideally,  $R\_ = R - RR$ . However, there could be a set of rules  $r$  in  $R\_$  from which one could derive or infer a restrictive rule in  $RR$ . So in reality,  $R\_ = R - (RR + RSE)$ , where  $RSE$  is the set of non-restrictive rules that are removed as side effects of the sanitization process to avoid recovery of  $RR$ ". It is shown below in figure:

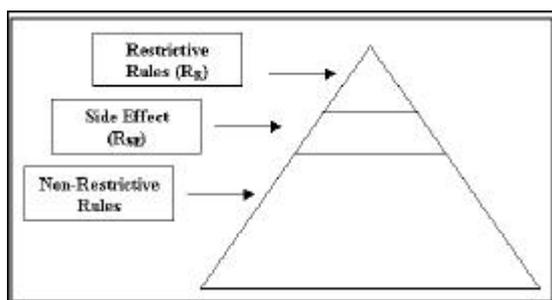


Figure 1: The inference problem in association rule-mining.

The security impact of DM is analyzed in [6] and some possible approaches to the problem of inference and discovery of sensitive knowledge in a data mining context are suggested. The proposed strategies include fuzzyfying and augmenting the source database and also limiting the access to the source database by releasing only samples of the original data. Clifton [7] adopts the last approach as he studies the correlation between the amount of released data and the significance of the patterns which are discovered. He also shows how to determine the sample size in such a way that data mining tools cannot obtain reliable results.

Clifton and Marks in [6] also recognize the necessity of analyzing the various data mining algorithms in order to increase the efficiency of any adopted strategy that deals with disclosure limitation of sensitive data and knowledge. The solution proposed by Clifton in [7] is independent from any specific data mining technique; other researchers [8] propose solutions that prevent disclosure of confidential information for specific data mining algorithms such as association rule mining and classification rule mining.

Classification mining algorithms may use sensitive data to rank objects; each group of objects has a description given by a combination of non sensitive attributes. The sets of descriptions, obtained for a certain value of the sensitive attribute, are referred to as description space. For Decision-

Region-based algorithms, the description space generated by each value of the sensitive attribute can be determined a priori. The authors in [4] first identify two major criteria which can be used to assess the output of a classification inference system and then they use these criteria, in the context of Decision-Region based algorithms, to inspect and to modify, if necessary, the description of a sensitive object so that they can be sure that it is not sensitive.

There is a large amount of work related to association rule hiding. Maximum researchers have worked on the basis of reducing the support and confidence of sensitive association rules ([1, 2, and 5]). ISL and DSR are the common approaches used to hide the sensitive rules. Actually any given specific rules to be hidden, many approaches for hiding association, classification and clustering rules have been proposed. Some of the researchers have used data perturbation techniques ([3]) to modify the confidential data values in such a way that the approximate data mining results could be obtained from the modified version of the database. Some researchers also recognize the necessity of analyzing the various data mining algorithms in order to increase the efficiency of any adopted strategy that deals with disclosure limitation of sensitive data and knowledge. Also disclosure limitation of sensitive knowledge by data mining algorithms, based on the retrieval of association rules, has been recently investigated. The proposed work also has the basis of reduction of support and confidence of sensitive rules in which some modified terms and some new variable are used to do the job.

### 3. Problem Statement

The problem of sensitive rule hiding is described as follows: *Given a transaction database,  $MST$ ,  $MCT$ , a set of strong rules, and a set of sensitive items, how can we modify the database such that using the same  $MST$  and  $MCT$ , the set of strong rules in the modified database satisfies all the constraints: 1) no sensitive rule, 2) no lost rule, and 3) no false rule?*

Let  $D$  be the database of transactions and  $J = \{J_1, \dots, J_n\}$  be the set of items. A transaction  $T$  includes one or more items in  $J$ . An association rule has the form  $X \Rightarrow Y$ , where  $X$  and  $Y$  are non-empty sets of items (i.e.  $X$  and  $Y$  are subsets of  $J$ ) such that  $X \cap Y = \text{Null}$ . A set of items is called an itemset, while  $X$  is called the antecedent. The support of an item (or

itemset)  $x$  is the percentage of transactions from  $D$  in which that item or itemset occurs in the database The confidence or strength  $c$  for an association rule  $X \Rightarrow Y$  is the ratio of the number of transactions that contain  $X$  or  $Y$  to the number of transactions that contain  $X$ .

The problem of mining association rule is to find all rules that have support and confidence greater than user specified minimum support threshold (MST) and minimum confidence threshold (MCT).

#### 4. Proposed Approach

Internet communication technology has made this world very competitive. In their struggle to keep customers, to approach new customers or even to enhance services and decision making, data owners need to share their data for a common good. Privacy concerns have been influencing data owners and preventing them from achieving the maximum benefit of data sharing. Data owners usually sanitize their data and try to block as many inference channels as possible to prevent other parties from finding what they consider sensitive. Data sanitization is defined as the process of making sensitive information in non-production databases safe for wider visibility. However, sanitized databases are presumed secure and useful for data mining, in particular, for extracting association rules.

To hide any specified association rule  $X Y$  this algorithm works on the basis of confidence ( $X Y$ ) and support ( $X Y$ ). To hide any sensitive rule  $X Y$ , this algorithm first finds the value of support (sup) and confidence (conf) in the available set of rules and then it computes the support and confidence of the sensitive rule using following:

$$\text{Confidence (X Y)} = (\text{conf} * 1/3);$$

$$\text{Support (X Y)} = (\text{sup} * 1/3);$$

##### Procedure:

```
//find value of support and confidence Select confidence into
conf from database. Select support into supp from database.
For each X
{
//Now check all the rules containing sensitive element x. For
each rule R which contain X on LHS OR RHS.
{
While (conf(R) >= MCT)
{
Set confidence(X Y) = (conf * 1/3); Set support (X Y) = (sup
* 1/3);
}
}
}
End of procedure
```

#### 5. Result Analysis & Conclusion

##### An Example Data Set

Suppose there is a database of transactions as below:

TID	Items
T1	ABC
T2	B
T3	ACD
T4	AB
T5	ABD

One has also given a MST of 60% and a MCT of 70%. One can see four association rules can be found as below:

- A B (60%, 75%)
  - B A (60%, 75%)
  - A D (60%, 75%)
  - D A (60%, 100%)
- Now there is a need to hide A.

To hide A. After 1 pass the status of database is as follows:

	support , confidence	
A B	20%	25%
B A	50%	60%
A D	20%	25%
D A	43%	60%

So it is clear that this approach is hiding all the given sensitive rules successfully without any side effect in small as well as large databases. Previous algorithm (Hiding counter) is also hiding all sensitive rules without any side effect in small datasets. But when the transactions are too many then in that case this algorithm is having a major problem. It takes too many passes to bring down the support below the threshold. For example: Suppose a database has 100 transactions. If support of a rule says  $X \rightarrow Y$  is 80%. If MST is 30% then this algorithm will require 170 passes to reduce the support to 29%. But our proposed algorithm will take 1 pass to reduce the support below 30%. So it is clear that the proposed algorithm is more efficient.

#### References

- [1] Shyue-Liang Wang, Yu-Huei Lee, Steven Billis, Ayat Jafari "Hiding Sensitive Items in Privacy Preserving Association Rule Mining" 2004 IEEE International Conference on Systems, Man and Cybernetics
- [2] Vassilios S. Verykios, Ahmed K. Elmagarmid, Elisa Bertino, Yucel Saygin and Elena Dasseni" Association Rule Hiding", IEEE Transactions on Knowledge and Data Engineering, Vol. 16No. 4, April 2004.
- [3] R. Agrawal and R. Srikant, "Privacy preserving data mining", In ACM SIGMOD Conference on Management of Data, pages 439450, Dallas, Texas, May 2000.
- [4] Hung Wu, Chia-Ming Chiang, and Arbee L.P. Chen, Senior Member, IEEE Computer Society Hiding Sensitive Association Rules with Limited Side Effects IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 19, NO.1, JANUARY 200
- [5] S. Oliveira, o. Zaiane, "Algorithms for Balancing Privacy and Knowledge Discovery in Association Rule Mining", Proceedings of 71 th International Database Engineering and Applications SYmposium (IDEAS03), Hong Kong, July 2003.
- [6] C. Clifton and D. Marks, "Security and Privacy Implications of Data Mining," Proc. 1996 ACM

- Workshop Data Mining and Knowledge Discovery, 1996.
- [7] C. Clifton, "Protecting against Data Mining through Samples," Proc. 13th IFIP WG11.3 Conf. Database Security, 1999.
- [8] T. Johnsten and V.V. Raghavan, "Impact of Decision-Region Based Classification Mining Algorithms on Database Security," Proc. 13th IFIP WG11.3 Conf. Database Security, 1999.