International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 5, May - 2014

# Vernam Cipher Encrypted Approach for Classification Rule Hiding to Preserve the Privacy in Database

## Encryption approach to Preserve Privacy in Database

Deepak Antil1, Aakanksha Mahajan2
Department of Computer Science Engineering
Panipat Institute of Engineering and Technology
Samalkha (Panipat) India

***Abstract*— sensitive frequent pattern or classification rule hiding is an important issue in privacy preserving in database. In this era of information explosion and rapid development of the Internet, the data stored in the database is usually continuously updated. Existing frequent pattern hiding algorithms gradually become inadequate because those algorithms are originally designed for static database and thus they cannot handle incremental datasets effectively and efficiently. It's common to share data information between two organizations in different application. When information are to be shared between organizations, there could be some important patterns which not be disclosed to the other organizations. We notice such type of problem as sensitive classification rule hiding. We propose an approach for sensitive classification rule hiding. 1st we find the helping transactions of sensitive rules. Then we replace sensitive values with vernam cipher Encrypted values to hide a given sensitive classification rules.**

***Keywords—Classification Rule Hiding, Encrypted, Data Mining, Vernam cipher, Privacy Preserving.***



### I.INTRODUCTION

Over the past few years, there has been a tremendous growth in the amount of private data collected about individuals that can be collected and analyzed. This data comes from a variety of sources including medical, financial, library, telephone, and shopping records. With the rapid growth in database, networking, and computing technologies, such data can be integrated and analyzed digitally. On the one hand, this has led to the development of data mining tools that aim to infer useful trends from this data. But, on the other hand, easy access to personal data poses a threat to individual privacy [1]. With it new threats to privacy of the data are also increases. An interesting way data mining research has been developed, known as privacy preserving data mining. Main objective of these algorithms is the extraction of relevant data from database, As well as protecting private information simultaneously. The main objective in privacy preserving data mining (PPDM) is to develop algorithms for modifying the original data in such a way, so that the private data and knowledge remain private even after the mining process. PPDM approaches can be classified as two categories. The first is related to the data parse and is known as data hiding, while the second concerns the information ,or else the knowledge, that a data mining method may discover after having analyzed the data , and is known as knowledge hiding. Data hiding tries to remove confidential or private information from the data before its disclosure.

Here we only focus on privacy preserving approaches [2] that can preserve privacy on any sensitive knowledge patterns. These approaches modify the original data in such a way that some sensitive knowledge patterns are hidden on mining the data. In this paper, we focus on the problem classification rules privacy preservation. Classification rule hiding algorithms consider a set of classification rule as a sensitive and aim to protect them. Our goal is the successful hiding of the sensitive classification rules.

In section II we present theoretical background and related work. In section III we define the problem. In section IV proposed algorithm, Experimental results are in section V. Conclusion in section VI. Future work in VII section, References in next section.

### II. THEORETICAL BACKGROUND AND RELATED WORK

#### *A. Privacy preserving techniques*

Main objective in privacy preserving is to protect the sensitive data before it is used for analysis. However the data may be distributed or non-distributed. In such type of problem appropriate algorithms or techniques should be used which preserves sensitive information in the knowledge discovery process. There are many approaches adopted for addressing the privacy preserving data mining.

*a. Secure computation*

The history of the multi-party computation problem is extensive since it was introduced in Greedy Approach [3] and extended by Unknowns to Prevent Discovery of Association Rules [4] and many others. The basic 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. Security sum is a very simple and useful method which base on secure multiparty computation. It is used for get the sum of data from the different site.

*b. Distribution of data*

On behalf of distribution of data, the PPDM algorithms can be divided into two sub categories, centralized and distributed data. In centralized database data are all stored in a single database; while, in a distributed database data are stored in different sites. Distributed data can be further subdivided into horizontal and vertical data distribution.

*c. Modification of data*

Data modification is used to ensure high privacy protection when it is necessary to modify the original values. Methods of modification include:

- Merging or aggregation combination of several values.
- Replacement of an existing attributes value with a Encrypted value.
- Interchanging, swapping the values.
- Data mining algorithm

The most important ideas have been developed for classification data mining algorithms, like decision tree inducers, association rule mining algorithms, clustering algorithms and Bayesian networks.

*d. Rule hiding or Data hiding*

The PPDM [5], algorithms can be further divided into two types, rule hiding and data hiding according to the purposes of hiding. In rule hiding, we remove the sensitive rule defined from original database after applying data mining algorithms. Data hiding where the sensitive data from original database like name subject and branch that can be linked, directly or indirectly, to an individual person may hided.

*e. Privacy preservation*

This refers to the privacy preservation technique used for the selected data modification. The techniques are:

- Hiding Knowledge in XML Document Collections

Identify Sensitive Knowledge
Identify Appropriate Classes of Data Mining Algorithms
Formulate Security Policies
Sanitize Data
Report Generation

- A Border-Based Approach for Hiding Sensitive Frequent Item sets.
- Heuristic-based techniques modify selected values
- Cryptography-based techniques

Hiding of classification rules is our main objective. Over the last few years interest has increased towards dealing with the problem of hiding sensitive patterns and sensitive data. In a reconstruction algorithm [6] is proposed for classification rules hiding. They proposed an algorithm to preserve the privacy of the classification rules by using reconstruction technique. In which, only non-sensitive rules of the dataset are used to build a decision tree. Finally, the new dataset which contains only non-sensitive classification rules is reconstructed from the tree.

A data reduction approach was adopted in [7].They addressed the problem of sensitive classification rule hiding by using data reduction approach. i.e. removing the whole selected tuples in the given dataset.

In [8] data perturbation approach was proposed for hiding sensitive classification rules. This approach modifies the tuples of sensitive rules of a dataset Ds. In data perturbation approach tuples belonging to the sensitive rules are assigned to the non-sensitive rules based on their rank in the rule set.

In first step data reduction approach identifies the sensitive and non-sensitive rules, then it selects the tuples of the non-sensitive rules and assigns them to the perturbed dataset D1. It then proceeds to the sensitive rules. In each tuples of a sensitive rule some attribute-value are being changed in order to match a non-sensitive rule.

My approach for classification rule hiding a encryption based approach proposed in [9] and [10] [11] for association rules privacy preserving. They increase or decrease the support of item by placing encrypted values. It is difficult to know the value behind encrypted values. In [11] blocking based approach to preserve privacy for sensitive association rules in database. For hiding the sensitive rules, this algorithm replace the 1's or 0's by unknowns values or symbols to decrease their confidence in selected transaction. For decreasing the confidence of specified rules, first algorithm increases the support of rule on one side, while another decreases the support of rule another side. They can hide many rules at a time in rule set.

## III. PROBLEM DEFINITION

In some approaches, values are changed, values are replaced by some symbols for Example 1 is replaced by 0 and 0 is replaced by 1 and 0 is replace by $ or 1 is replace by @.Sometimes it may have bad consequences. Consider a research that will publish some of its data, and data is sanitized by replacing actual values by true or false values. Researchers may use this data in worst case, such wrong rules could be used for critical purposes. Therefore for many situations it is safer if the sanitization process place unknown values. This obscures the sensitive rules, while protecting the user of the data from learning 'false' rules.

### *A. Problem Statement*

Given a dataset D, a class attribute C, a set of classification rules R over D, as well as a sensitive rule Rs , we want to find a dataset D' such that when mining D' for classification rules using the same parameters as those used in the mining of D, only the (nonsensitive) rules in R −Rs can be derived.

## IV. ALGORITHM

This algorithm applies where we can store encrypted values for some attributes, when actual values are not available or confidential. Here we propose encryption based algorithm to preserve privacy for sensitive classification rules in database. To hide sensitive rules, proposed algorithm replaces the A's or B's by unknown's encrypted values in selected transactions. Main objective of the algorithm here is to obscure a given set of sensitive rule by replacing known values with encrypted values.

In this algorithm, for each sensitive rule, it reads the original database and find outs the transactions supporting sensitive rules. We can say transaction supports any rule when the left side of the rule (attribute –value) pair is a subset of attribute values pair of the transaction and the right hand side of the rule is same as the class attribute of the transaction. Then for each transaction that supports sensitive rule, algorithm places encrypted values in place of attribute value which appears in rule. This procedure continues until all the sensitive rules are hidden. Finally the sanitized dataset which contains encrypted values is released any one.

Example:

A college dataset as a example transaction database D is shown in table 1.In our example name, designation, department, location and subject are the attributes of the transaction. And Department is the class attribute.

Fig. 1 contains the set of classification rules generated from this dataset. Now suppose rule 3 is considered as sensitive rule. We have needed to find out the set of transactions that satisfies rule 3. We can find out that tuples no. 7 and 11 are the supporting transactions of rule 3. So for hiding rule 3, we place the encrypted values in place of sensitive rules in transaction. The modified dataset is shown in table 2. Fig. 2 shows the classification rules generated from the sanitized dataset.

Algorithm

Input: Initial database D and set of classification rules R.
Step1. Begin
Step2. *D* ⇦ {}
Step3. *D1*⇦ *D*
Step4. *Rs* ⇦ set of sensitive rules.
Step5. For each rule *Ri* ∈ *Rs*
Step6. {
Step7. For each tuple t ∈ D1 do
Step8. {
Step9. If *t* supports *Ri* then
Step10. *Ti* ⇦ *t*. (add *t* to *Ti*)
Step11. }
Step12. }
Step13. For each *Ti*, where *Ri* ∈ *Rs* do
Step14. {
Step15. For each transaction *t* ∈ *Ti* do
Step16. {
Step17.(I) Treat each plain text alphabet as a number in an increasing sequence ,i.e. A=1,B=2……..Z=26.
(ii) Select key k=1, 2…25.
(iii)Do the same or each character of the input cipher text.
(iv)Add each number k corresponding to the plain text alphabet to the corresponding input cipher text alphabet number.
(v)If the sum produced is greater than 26, subtract 26 from it.
(vi)Translate each number of the sum back to the corresponding alphabet. This gives the output cipher text.
Step18. Update *D1*.
Step19. }
Step20. }
Step21. *D* ⇦ *D1*
Step22. End
Output: sanitized database D

Table1.Dataset

| NAME | DESIGNATION | DEPARTMENT | LOCATION | SUBJECT |
|---|---|---|---|---|
| NEETU | ASSISTANT PROFESSOR | CIVIL | DELHI | CONSTRUCTION |
| DEEPAK | PROFESSOR | CSE | SONIPAT | JAVA |
| SWEETY | PROFESSOR | CSE | SONIPAT | DBMS |
| AAKANKSHA | HEAD OF DEPARTMENT | ME | KARNAL | MACHIN DESIGN |
| NEETU | PROFESSOR | CIVIL | KARNAL | CONSTRUCTION |
| ANCHAL | PROFESSOR | CSE | PANIPAT | DS |
| NAVNEET | ASSISTANT PROFESSOR | ECE | PANIPAT | ROBOTICS |
| SWEETY | PROFESSOR | ME | DELHI | FLUID |
| DEEPAK | ASSISTANT PROFESSOR | CSE | DELHI | JAVA |
| VIKRAM | HEAD OF DEPARTMENT | CSE | PANIPAT | JAVA |
| NAVNEET | PROFESSOR | ECE | SONIPAT | ROBOTICS |
| VIKAS | ASSISTANT PROFESSOR | CIVIL | KARNAL | CONSTRUCTION |

1. (NAME=NEETU) and (SUBJECT=CONS-TRUCTION) =>DEPARTMENT=CIVIL
2. (NAME=DEEPAK) and (SUBJECT= JAVA) =>DEPARTMENT=CSE
3. (NAME=NAVNEET) and (SUBJECT= ROBOTICS) =>DEPARTMENT=ECE

Fig.1. Classification Rules

Key = JJJJJJJ, JJJ, JJJJJJJJ



| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | N | A | V | N | E | E | T | | R | O | B | O | T | I | C | S | | E | C | E |
| PLAIN TEXT – | | 13 | 0 | 21 | 13 | 4 | 4 | 19 | | 17 | 14 | 1 | 14 | 19 | 8 | 2 | 18 | | 4 | 2 | 4 |
| + | | | | | | | | | | | | | | | | | | | | | |
| KEY | | 10 | 10 | 10 | 10 | 10 | 10 | 10 | | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | | 10 | 10 | 10 |
| TOTAL | | 23 | 10 | **31** | 23 | 14 | 14 | **29** | | **27** | 24 | 11 | 24 | **29** | 18 | 12 | **28** | | 14 | 12 | 14 |
| SUB IF >=26 | | 23 | 10 | 5 | 23 | 14 | 14 | 3 | | 1 | 24 | 11 | 24 | 3 | 18 | 12 | 2 | | 14 | 12 | 14 |
| CIPHER TEXT | | W | J | E | W | N | N | C | | A | X | K | X | C | R | L | B | | N | L | N |

Table 2. Sanitized dataset

| NAME | DESIGNATION | DEPARTMENT | LOCATION | SUBJECT |
|---|---|---|---|---|
| NEETU | ASSISTANT PROFESSOR | CIVIL | DELHI | CONSTRUCTION |
| DEEPAK | PROFESSOR | CSE | SONIPAT | JAVA |
| SWEETY | PROFESSOR | CSE | SONIPAT | DBMS |
| AAKANKSHA | HEAD OF DEPARTMENT | ME | KARNAL | MACHIN DESIGN |
| NEETU | PROFESSOR | CIVIL | KARNAL | CONSTRUCTION |
| ANCHAL | PROFESSOR | CSE | PANIPAT | DS |
| **WJEWNNC** | ASSISTANT PROFESSOR | **NLN** | PANIPAT | **AXKXCRLB** |
| SWEETY | PROFESSOR | ME | DELHI | FLUID |
| DEEPAK | ASSISTANT PROFESSOR | CSE | DELHI | JAVA |
| VIKRAM | HEAD OF DEPARTMENT | CSE | PANIPAT | JAVA |
| **WJEWNNC** | PROFESSOR | **NLN** | SONIPAT | **AXKXCRLB** |
| VIKAS | ASSISTANT PROFESSOR | CIVIL | KARNAL | CONSTRUCTION |

1. (NAME=NEETU) and (SUBJECT=CONS-TRUCTION) =>DEPARTMENT=CIVIL
2. (NAME=DEEPAK) and (SUBJECT= JAVA) =>DEPARTMENT=CSE

Fig 2. Sanitized dataset's classification rule



## V. EXPERIMENTS AND RESULTS

In our experiment we have used two dataset. The detail of each dataset is

TABLE3. DETAIL DATASET

| SET | INSTANCE | ATTRIBUTES | RULE |
|---|---|---|---|
| **SELLER** | 40 | 4 | 3 |
| **SPORTS** | 20 | 3 | 5 |

The experiments have two parts. From the set of classification rules one rule is randomly selected as the sensitive rule. Then a sensitive rule is randomly selected for hiding. Proposed algorithm is applied to hide the sensitive rule. Another experiment is for hiding more than one rule. The procedure is same as single rule hiding. We can check the algorithm for multi-rule hiding.

### *A. Evaluation*

#### *a. Hiding Failure and Side effects*

First evaluation is the hiding failure, i.e. the percentage of sensitive information that is still discovered, after the data has been sanitized, gives an estimate of the hiding failure parameter. Here, we are placing encrypted values in transaction which satisfy the sensitive classification rule such that they no longer satisfy sensitive rule. So, sensitive rule will not discovered from sanitized dataset.

Side effects generated due to the rule hiding. We can find out the side effects in terms of number of false drop rules and number of ghost rule. False drop rules are nonsensitive rules which were present in the original dataset but hidden from the sanitized dataset. Ghost rules are those rules which were not in the original dataset but are present in the sanitized dataset.

#### *b. Experiment result*

Table 3 shows experiment results when a sensitive rule is hidden. Here, nonsensitive rule is discovered from the sanitized dataset. Table 4 shows the experiment results when more than one rule are hidden. From table 3 our proposed approach can be used to hide sensitive classification rule with minimum side effects.

EXPERIMENTAL RESULTS (MULTIPLE RULES)

Table 4.

| SET | SENSITIVE RULE | GHOST RULE | DROP RULE | HIDDEN RULE |
|---|---|---|---|---|
| **SELLER** | 0 | 0 | 0 | 2 |
| **SPORTS** | 0 | 0 | 0 | 3 |

## VI. CONCLUSION

In this paper, we have proposed an algorithm to preserve privacy of sensitive classification rules. The algorithm places encrypted values in place of known values in the transactions that support the sensitive rules. So, from the sanitized dataset's the sensitive rules are no longer generated.

## VII. FUTURE WORK

It's not so much difficult to find out the plain text value from the cipher text values in vernam cipher .In future, the algorithm can be modified such that we can use more secure algorithm like AES, RSA etc. or we can insert some dummy transactions in the dataset in place of sensitive rules.



## REFERENCES

[1] "*An Efficient method for knowledge Hiding Through database Extension*" B.Murugeshwari Research Scholar Anna University Chennai, India, Dr.K.Sarukesi Vice Chancellor Hindustan University Chennai, India, Dr.C.Jayakumar Head of the Dept in CSE Eswari Engg College Chennai, India,2011

[2] "*Privacy Preserving Based on Association Rule Mining*" Tinghuai Ma, Sainan Wang School of Computer & Software Nanjing University of Information Science & Technology Nanjing, China thma@nuist.edu.cn ZhongLiu Sichuan College of Architectural Technology Deyang, China liuzhong@scac.edu.cn 2010

[3] "*Privacy Preserving Association Rules by Using Greedy Approach*" Chieh-Ming Wu, Yin-Fu Huang and Jian-Ying Chen Graduate School of Engineering Science and Technology National Yunlin University huangyf@el.yuntech.edu.tw IEEE 2008

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

[5] "*Full-scale Privacy Preserving for Association Rule Mining*" Tinghuai Ma, Jiazhao Leng School of Computer & Software Nanjing University of Information Science & Technology Nanjing, China thma@nuist.edu.cn Keyi Li Department of Automation Beijing Forestry University Beijing, China 2010

[6] "*A Reconstruction-Based Algorithm For Classification Rules Hiding*" Natwichai, J., Li, X., Orlowska, M.E.Proceedings 17th Australasian Database Conference, pp. 49–58, 2006.

[7] "*Data Reduction Approach for Sensitive Associative Classification Rule Hiding*", J. Natwichai, X. Sun, X. Li, 19th Australian Database Conference (ADC2008), Wollongong, Australia, 2008

[8] "*A Data Perturbation Approach to Sensitive Classification Rule Hiding*" Aggelos Delis, Vassilios S. Verykios, SAC'10, ACM, Sierre, Switzerland, March 22-26, 2010.

[9] "*Privacy Preserving Association Rule Mining*" Y. Saygin, V.S. Verykios, A.K. Elmagarmid. Proceedings International Workshop on Research Issues in Data Engineering (RIDE 2002), pp.151–163, 2002.

[10] "*A Border-Based Approach for Hiding Sensitive Frequent Itemsets*" Xingzhi Sun in their research paper, Fifth IEEE International Conference on Data Mining (ICDM'05) 1550-4786/05 © 2005.

[11] "*Blocking based approach for classification Rule hiding to Preserve the Privacy in Database*" Anita A. Parmar,Udai Pratap Rao, Dhiren R. Pate, 2011 International Symposium on Computer Science and Society, 978-0-7695-4443-4/11©2011 IEEE