Sensitive Information Hiding with Data Mining

Apoorva Patil (W1188522), Joyce Annie George (W1184679), Krima Shah (W1183143),Prajakta Deosthali (W1185340), Sasikumar Ravichandran (W1182978), Vinod Kumar Sivalingam (W1184028)

Computer Science Engineering
Department Santa Clara University
Santa Clara, CA, USA

Abstract—Web Data mining has attracted a great deal of attention in information industry and in society as a whole in recent years, due to the wide availability of huge amount of data and the imminent need for such data to be converted into useful information and knowledge. The real privacy concerns are with sensitive data which needs to be protected from unauthorized access. In this research paper, we present an overview of rapidly emerging research areas in privacy preserving of data mining. The two commonly used techniques in information security hiding are Association rule mining and decision tree. This paper compares and analyses the listed data mining techniques and current studies of Private Preserving Data Mining (PPDM) which mainly focus on how to reduce the privacy risk brought by data mining.

.

1. Introduction

Data mining is a process of automatic or semi-automatic analysis of large quantities of data to extract previously unknown, interesting patterns. It is one of the important and fundamental steps in 'Knowledge Discovery in Database' or KDD. Data mining deals with data from varied sources, for example, financial transactions, healthcare records. Such data contains private and sensitive information of millions of users. Extracting useful information from such datasets is of paramount importance to the businesses and other stakeholders like governments, health care agencies, law enforcement agencies etc. At the same time, the easy access to this sensitive data poses a great security threat. This data in wrong hands can lead to serious threats to personal security. To address this threat and growing concern among users, many different techniques have been designed to analyze and access sensitive and private data.

One of the most well-known techniques is Privacy Preserving Data Mining. As the name suggests, this methodology ensures the privacy of the date being mined.

A set of approaches are classified based on following dimensions:

1. Data distribution
2. Data modification
3. Data mining algorithm
4. Data or rule hiding
5. Privacy preservation

Based on above approaches, following techniques are devised over the years.

1. Association Rule
2. Decision Tree
3. Random Value perturbation
4. Privacy Integrated Queries (PINQ)
5. UML diagrams

The main focus of this research paper is on first two methods mentioned. viz., Association Rule and Decision Tree. For the sake of completeness, we will explain the other techniques in brief.

Random Value Perturbation - It hides the sensitive data, but still allows to estimate the underlying distribution in dataset. The sensitive values are modified and hence successfully hidden using a randomized process. Two common ways of randomization are Value-Class Membership and Value Distortion.

Privacy Integrated Queries - It provides differential privacy. This allows analysts' access to records in the dataset through SQL like declarative language. It has additional features like grouping and joining records on sensitive attributes, analysis of text and unstructured binary data, modular algorithm design.

UML methodology - In this case, the privacy model is depicted with the help of different diagrams like logical diagrams, use case diagrams, collaboration, distribution as well as activity and scenario diagrams.

Data mining techniques are being widely in the analysis of large datasets. It is used in the field of Marketing, Finance, Business, etc. Though being helpful, these tools might pose a threat to privacy and data security. Sensitive user information like credit card number, social security number, and medical history can often be misused. This poses a need to develop some mechanisms to ensure data privacy and security of sensitive information. The technique used for security information hiding is called Privacy Preserving Data Mining (PPDM). The basic idea of PPDM is to modify the data in such a way so as to perform data mining algorithms effectively without compromising the security of sensitive information contained in the data.

The statistical approach has been used to extract the facts from association rules, decision trees and clustering. These approach are very popular because of its high performance. The intent of privacy preserving data mining is to hide or trim sensitive information so that the data cannot be through data mining techniques.

Privacy preserving data mining (PPDM) is a novel study area that examines the troubles which occurs after applying the data mining techniques. PPDM can be summarized as below:

1. The data is changed before delivering it to the data miner. 2. The data is circulated between two or more locations which work together using a protocol to study data mining results without

disclosing any information about the data at their separate sites. 3. While using a demonstration to categorize data, the classification results are only shown to the selected party, who do not learn anything, but can check for availability of certain rules without revealing the rules.

2. Association Rule Hiding:

Association Rule Hiding is a PPDM method which use Association Rule Mining method that identifies interesting correlation or association in the huge volume of relations. Classifying associative rules of a transactional database in data mining may reveal the privacy and confidentiality of an individual and organization. This problem is resolved using Association Rule Hiding (ARH) techniques in Privacy Preserving Data Mining (PPDM). We examine the privacy issues of a wide group of rules called the association rules. This study on Association Rule Hiding method in data mining accomplishes the generation of sensitive association rules by the way of hiding. The sensitive data is either trimmed or modified and therefore the data mining algorithm could find only the non-sensitive data in the database. The idea used in all the algorithms is to lessen the confidence or support of the sensitive rules.

2.1. Encrypt Decrypt Module

The client/owner encrypts the data using encrypt/decrypt (E/D) module which is measured as a black box from its perspective. Encrypt/decrypt (E/D) module is used for changing the input data into an encrypted database. On the contrary, the server accomplishes data mining operations and transferred the patterns in the encrypted form to the owner of the data. The encryption scheme encompasses property where the revisited supports are not true supports. The E/D module reclaims the true individuality of the true supports and returned patterns. An encryption scheme is called as Rob Frugal. It is used to alter the client data before it send to the server.

2.2. Problem Formulation:

Consider a set of dissimilar literals L = {l_1, l2, …., lm } called items. For a given set of transactions D and each transaction ‘T’ is a set of items which is given as T ⊆ L. Association rule is an inference of the form A−>B where each is associated with the set of literals A ⊂ L ,B ⊂ L and A∩B = ϕ. A and B are called as antecedent/body and consequent/head of the rule respectively.

The strength of the rule to check whether it is calculated using the following limitations called confidence and support of the rule. These limitations help in determining the interestingness of a rule. For a given rule A⇒B, support is the proportion of transaction that contains both A and B (A U B) or is the percentage of transactions jointly covered by the RHS and LHS and is calculated as:

$$\mathbf{S = |A\ U\ B\ | / |N|}$$

Where, N is the number of transactions.

Confidence is the proportion for a transaction that contains the proportion of transactions covered by the LHS that are also covered by the RHS and is calculated as

$$\mathbf{Z = |A\ U\ B\ | / |A|}$$

For the database given in Table1, with a least support of 33% and minimum confidence 70% resulting nine association rules could be found:

Z => X (66.667%, 100%),X ,Y => Z (50%, 75%),
Y =>Z ,X (50%, 75%),Z ,Y => X(50%, 100%)
Z => X,Y (50%, 75%),Z ,X =>Y (50%, 75%)
Y => Z(50%, 75%),Z => Y(50%, 75%)
Y => A(66.667%, 100%)

TID	ITEMS
T1	XYZ
T2	XYZ
T3	XYZ
T4	XY
T5	X
T6	XZ

Table 1

In General, number of association rules exposed in a given database is very huge. It is witnessed that a considerable proportion of these rules are redundant and unworkable. A user should be offered only with the ones which are unique, different and fascinating. To report this issue, introduced a notion for brief representation of association rules, called representative rules (RR). RR is a least set of rules that allow inferring all association rules without granting access to a database. In a perception of cover operator was presented for a set of association rules from a given association rule.

The main objective of this is to suggest a new strategy where sensitive data can be avoided. Data should be manipulated or mislead in such a way that sensitive information cannot be found through data mining techniques. While handling sensitive information it becomes very important to protect data against unauthorized access. The main problem encountered is the need to balance the confidentiality of the disclosed data with the authentic needs of the data users. The proposed algorithm are based on changing the database transaction, so that, the assurance of the rules can be reduced.

Hiding sensitive rules by changing the support and the assurance of the association rule or frequent item set as data mining mainly deals with generation of association rules. Most of the work done by data miner revolves around association rules and their generation. As it is known that association rule is a significant factor, which may cause harm to the private information of any business, defense or organization and raises the need of hiding this information.

2.3 Proposed Algorithm

In order to preserve an association rule, a new concept of ‘not altering the support’ of the sensitive items has been recommended in this paper. Based on this approach an algorithm has been suggested that a set of sensitive item(s) is passed and the proposed algorithm distorts the original database such that sensitive rules cannot be found through Association Rule Mining algorithms.

Input to suggested algorithm is a database, min_supp, min_conf, and a sensitive item(s) H (each delicate item in H is denoted by *h*) to be hidden and the aim is to mislead the database D such that no association rules enclosing *h* belongs to H either on the right or on the left can be found.

Input
(1) A source database D
(2) A minimum support.
(3) A minimum confidence.
(4) A set of sensitive items H.

Output a transformed database D’ where guidelines having H on RHS/LHS will be hidden.

```
1. Find all large item sets from D;
2. For all sensitive item h∈H {
3. If h is not a large item set then H=H- {h};
4. If H is empty then EXIT;
5. Select all the rules having h and store in U
6. Choose all the rules from U with h alone on LHS
7. Link RHS of selected rules and store in R;
8. Sort R in descending order by the number of reinforced items;
9. Choose a rule r from R
10. Calculate confidence of rule r.
11. If conf >min_conf then {
12. Find ={ t in D|t wholly supports r ;
13. If t has x and h then
14. Remove h from t
15. Else
16. Go to step 19
17. Find ={t in D|t does not support LHS(r) and moderately supports x;
18. Add h to t
19. Repeat
20. {
21. Select the first rule from R;
22. Calculate confidence of r ;
23. } Until(R is Null);
24. }
25. Update D with new transaction t;
26. Delete h from H;
27. Store U [i] in R;
28. i++, j++;
29. Go to step 7;
30. }
```

Output updated as the altered database proposed algorithm picks all the rules comprising sensitive item(s) either in the right or in the left. Then these rules are characterized in representative rules (RR) format. After this a rule from the set of RR’s, which has sensitive item on the left of the RR is selected. Now remove the sensitive item(s) from the transaction that wholly supports the RR i.e. it contained all the items in of RR selected and augment the same sensitive item to a transaction which partially supports RR i.e. where items in RR are present or only one of them is absent.
Like

Z -> X, Z-> Y, Can be represented as Z->XY

Now remove Z from a transaction where X, Y and Z are present and augment Z to a transaction where both of them (X and Y) are either present or only one of them is absent. Similar rules are being shaped containing sensitive item(s) are unseen.

Association rule mining is a significant data-mining task that finds exciting association among a huge set of data items. Later it may reveal patterns and numerous kinds of sensitive knowledge that are hard to find otherwise, it may pose a risk to the confidentiality of discovered confidential information. Such information is to be protected against illegal access. Aim of this work is to suggest a new strategy to avoid mining of sensitive data. Data should be distorted in such a way that sensitive information cannot be found through data mining techniques.

3. Decision Tree

A decision tree is a flowchart-like tree structure, where each node implies a test on an attribute value, each branch denotes an outcome of the test, and tree leaves signify classes or class distributions. Decision trees can easily be transformed to classification rules where it is basically a decision support tool which is used to make a predictive model. It works well on numerical as well as categorical data.

There are various decision tree algorithms, listed as follows

1. ID3 (Iterative Dichotomiser 3)
2. C4.5 (Successor of ID3)
3. CART (Classification and Regression Tree)
4. CHAID
5. MARS

The idea of using a decision tree ID3 algorithm, is to introduce a general privacy preserving data mining technique. Our goal is to create a background where each party has to disclose as little as possible, while still constructing a valid decision tree suitable for real-world applications. The aim is to create a valid decision tree based on minimum disclosure of sensitive information. First the data is modified by using different data modification and perturbation based approaches and then the decision tree mining algorithm is applied to clean dataset.

J. Ross Quinlan originally developed ID3 at the University of Sydney. He first presented ID3 in 1975 in his book titled ‘Machine Learning’. ID3 is based on the Concept Learning System (CLS) algorithm. ID3 is supervised learning algorithm which is used to builds tree based on the information obtained from the training occurrences and then uses the same to classify the test data. ID3 algorithm generally uses nominal attributes for classification with no missing values.

3.1 ID3 Algorithm

The key ideas behind the ID3 algorithm are:

(I) Each non-leaf node of a decision tree relates to an input attribute, and each arc to a possible value of that attribute. A leaf node matches to the expected value of the output attribute, when the input attributes are termed by the path from the root node to that leaf node.

(II) In a good decision tree, each non-leaf node should correspond to the input attribute which is the most informative about the output attribute amongst all the input attributes not yet considered in the path from the root node to that node. This is because we would like to predict the output attribute using the smallest possible number of questions on average.

(III) Entropy is used to regulate how informative a specific input attribute is about the output attribute for a subset of the training data. Entropy is a measure of uncertainty in communication systems.

3.1.1. Pseudo code of ID3 algorithm

```
1. ID3(Training data, Attributes)
2. node = DecisionTreeNode(Training data)
3.dictionary=summarizeExamples(Trainingdata,targetAttribute)
4. for key in dictionary
5. if dictionary[key] == total number of Training data
6. node.label = key
7. return node
8. if attributes is empty or number of Training data< minimum allowed per branch
9. node.label = most common value in Training data
10.return node
11. Best Attribute = the attribute with the most information gain
12. node.decision = Best Attribute
13. for each possible value of Best Attribute:
14.subset = the subset of training data that have value for Best attribute
15. if subset is not empty
16. node.addBranch(id3(subset,targetAttribute, attributes-Best Attribute
17.return node
```

Attributes is a list of features that may be tested by the learned decision tree. It returns a tree that correctly classifies the given examples. Assume that the target Attribute, which is the attribute whose value is to be predicted by the tree, is a class variable.

For the decision tree algorithm, ID3 was selected as it creates simple and effective tree with the smallest depth with multiple children and siblings. Multiple decision trees can be built from the similar training set by using the procedure described in the previous section, because of the undetermined selection criteria of the test attribute in the recursive case. The efficiency of a test element or attribute can be determined by its classification of the training set. A perfect attribute splits the outcomes as an exact classification, which achieves the goal of decision-tree classification. Iterative Dichotomiser 3 (ID3) selects the test attribute based on the information gain provided by the test outcome. Information Gain is the probable reduction in entropy caused by partitioning the examples according to a given attribute.

To unrealized the samples, we initialize both set of input sample dataset and perturbing dataset as empty sets, i.e. Unrealized training set is called. Consistent with the procedure described above, complete dataset is added as a parameter of the function because reusing pre-computed universal dataset is more efficient than recalculating universal dataset. The recursive function unrealized training-set takes one dataset in input sample dataset in a recursion without any distinctive requirement; it then updates perturbing dataset and set of output training data sets correspondent with the next recursion.

Therefore, it is evident that the unrealized training set process can be performed at any point during the sample gathering process.

```
Input: Unrealized training dataset
Output: Modified decision tree
If unrealized dataset is void then return default
Default ← minority –Value
Else
Tree ← best maximum value of information gain
Subtree← tree(root,best size)
Connect tree and subtree
Return tree
End
```

Similar to the traditional ID3 algorithm select attribute selects the test attribute using the ID3 criterion based on the information entropies, i.e., select the attribute with the greatest information gain. Algorithm Minority-Value regains the least recurrent value of the decision attribute, which performs the same function as algorithm Majority Value of the custom ID3 approach that is, getting the common frequent value of the decision attribute of TS. The decision attribute should be arbitrarily selected and compute the decision tree by calling the function Generate-Tree.

```
Attributes: set of attributes
Default: default value for the target predicate
Output: tree, a decision tree.

1. Best← Choose (attributes,size,( attribute, $T_s$ )
2. Tree← a new decision tree with root attribute best
3. Size← size of possible values ki in best
4. for each value vi of best do
5. $T'_i$←dataset in T' as best = $K_i$
6. $T'_p$←dataset in $T^p$ as best = $K_i$
7. subtree← Generate-tree
8. Link tree and subtree with a branch labelled k$i$
9. return tree
```

3.2. RIPPER Algorithm:

RIPPER algorithm to make the most of the information gain and increase the number of rules to covers the non negative rates. The RIPPER algorithm performs better than the ID3 algorithm. RIPPER algorithm performs two steps, adding the original rule R and after adding the condition R'or candidate rule measure the information gain (R, R') at true positives. Then it executes until the coverage of negative positive and negative true models in the data. Learning process, the training data is arranged by class labels in ascending order according to the corresponding class occurrences. Rules are then learned for the first m-1 classes, starting with the smallest one. Once a rule has been created, the instances covered by that rule are removed from the training data, and this is repeated until no instances from the target class are left. The algorithm then continues with the next class. Finally, when RIPPER finds no bonus rules to discover, a default rule is added for the last class.

In Ripper, conditions are added to the rule to maximize an information gain measure

$$Gain(R', R) = s.\,(log_2 \frac{N'_+}{N} - log_2 \frac{N_+}{N})$$

Where
R : The original rule
R' : The candidate rule after adding a condition
N (N'): the number of instances that are covered by R(R')
N_+ (N'_+): the number of true positives in R(R')

s: the number of true positives in R and R’ (after adding the condition)

Conditions are added to the rule until it covers no negative example.

$$rmv = \frac{p-n}{p+n} + 1$$

Where p and n: the number of true and false positives respectively.

Outer loop adds one rule at a time to the rule base and Inner loop adds one condition at a time to the current rule. The information gain measure is maximized by adding the conditions to the rule.

```
Ripper(Positive, Negative, n)
Rule Set ← LearnRuleSet(Positive, Negative)
For n times
RuleSet ← OptimizeRuleSet (RuleSet, Positive, Negative)
LearnRuleSet(Positive, Negative)
RuleSet is set to null
DL ← DescLen(RuleSet, Positive, Negative)
Repeat
Rule ← LearnRule(Positive, Negative)
Add Rule to RuleSet
DL’ ← DescLen(RuleSet, Positive, Negative)
If DL’ > DL
PruneRuleSet(RuleSet, Positive, Negative)
Return RuleSet
If DL1 <DL , DL ← DL’
Delete instances covered from Positive and Negative
Until Positive becomes Null
Return RuleSet
```

In order to overcome the issues of the prevailing privacy preservation approach recommend a new perturbation and randomization based approach that protects centralized sample data sets used for decision tree data mining. Privacy protection is applied to clean the samples prior to their release to third parties in order to lessen the threat of their inadvertent disclosure or theft.

4. Future Scope

The future direction of the security information hiding using association rule hiding techniques needs to handle the confidentiality of sensitive rules in terms of better data utility and optimal side effects on the modified transactional databases. As each user may have different concern over privacy, user-oriented privacy preserving techniques can be developed. Parallel algorithms could be developed to prevent revealing of sensitive association between items and to improve the performance of the algorithm for large and dynamic datasets. Most of the proposed research works are concentrating on side effects and numbers of sensitive rules are hidden from sanitized database. Those are not clearly stated about number of rules are hidden in each iteration, number of levels in multi-level sensitive rule hiding, number of scan needed for the database, computational efficiency in terms of memory and CPU time. In future, these objectives are also being considered and new techniques are to be proposed for hiding the sensitive association rules in privacy preserving data mining.

5. Conclusion

Association rule hiding technique can be effectively used to hide sensitive data without losing the information contained in it. There are no unexpected side effects of this technique as well.

ID3 algorithm selects the best attribute based on the concept of entropy and information gain for developing the tree. Decision trees are simply responding to a problem of discrimination. It is one of the few methods that can be presented quickly enough to a non-specialist audience. It allows to represent data processing without getting lost in complex mathematical formulations.

RIPPER algorithms which are very suitable for decision tree learning after completion of the unrealized dataset. RIPPER algorithm improvements have been created a rule learner and finally the results become unrealized dataset.

Privacy is of utmost importance to users. The analysis and understanding of data holds great potential for the stakeholders. The aim of various Privacy preserving data mining techniques is to find a golden mean. Based on the type of data, different techniques are used. More and more research in this area is being done to provide better solutions. In some cases, decision tree provides optimum result, whereas in some other association rule hiding is a better choice. The choice of technique solely depends on the type of data.

6. References

[1] V. S. Verykios,, Ahmed K. Elmagermld , Elina Bertino, Yucel Saygin, Elena Dasseni, “Association Rule Hiding,” IEEE Transactions on knowledge and data engineering, vol.6, no.4, (2004)

[2] Lu-Feng W:Association Rule Mining Algorithm Study and Improvement, In the 2nd International Conference on Software Technology & Engineering (ICSTE),362-364, (2010)

[3] Saygin Y., Verykios V.S. and Elmagarmid A.K., “Privacy preserving association rule mining,” IEEE Proceedings of the 12th Int’l Workshop on Research Issues in Data Engineering, pp. 151 – 158, (2002)

[4] Benny Pinkas ,HP Labs, Cryptographic Technique for Privacy PreservingDatamining,SIGKDD Explorations, Vol.4, Issue 2 ,2002

[5] S-L. Wang, Yu-Huei Lee, S.Billis and A. Jafari, “Hiding sensitive items in privacy preserving association rule mining,” IEEE International Conference on Systems, Man and Cybernetics, vol. 4, pp. 3239 – 3244, (2004)

[6] Marzena Kryszkiewicz. “Representative Association Rules”, In proceedings of PAKDD’98, Melbourne,Australia(Lecture notes in artificial Intelligence,LANI 1394, Springer-Verleg, pp 198-209, (1998)

[7] R. Agrawal and R. Srikant, “Privacy Preserving Data Mining,” Proc. ACM SIGMOD Conf. Management of Data (SIGMOD ’00), pp. 439-450, May 2000.

[8] Q. Ma and P. Deng, “Secure Multi-Party Protocols for Privacy Preserving Data Mining,” Proc. Third Int’l Conf. Wireless Algorithms,

Systems, and Applications (WASA ’08), pp. 526-537, 2008.

[8] Y. Zhu, L. Huang, W. Yang, D. Li, Y. Luo, and F. Dong, “Three New Approaches to Privacy-Preserving Add to Multiply Protocol and Its Application,” Proc. Second Int’l Workshop Knowledge Discovery and Data Mining, (WKDD ’09), pp. 554-558, 2009.