IJARIIT

INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X

Impact factor: 4.295

(Volume 4, Issue 3)

Available online at: www.ijariit.com

Protecting information by hiding sensitive data attributes

Sighila P
sighila@gmail.com
PSG College of Technology,
Coimbatore, Tamil Nadu

Sangeetha S
vns.sangeetha@gmail.com
PSG College of Technology,
Coimbatore, Tamil Nadu

ABSTRACT

Data mining aims at extracting hidden information from data. The process of discovering useful patterns and relationships in a large volume of data is called data mining. The goal of the data mining process is to extract information from a data set and transform it into an understandable structure. It involves data bases, data management aspects, visualization & online updating. Data mining poses a threat to information privacy. Privacy-preserving data mining hides the sensitive rules and prevents the data from being disclosed to the public.

The objective is to propose a novel association rule hiding (ARH) algorithm to hide the sensitive attributes. A function is used to obtain a prior weight for each transaction, by which the order of transactions modified can be efficiently decided. Apriori is used to find the frequent item set with minimum support and confidence. Sensitive rules are generated based on frequent item sets and FHSAR algorithm is used for hiding sensitive association rules.

This paper analyses the dataset obtained from SPMF an open source data mining library which is prepared based on real-life data. This paper shows the effectiveness of the algorithm.

Keywords: *Data mining, FHSAR. Data attributes.*

1. INTRODUCTION

Data mining is the computing process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics and database systems. It is an inter disciplinary subfield of computer science. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.

Privacy preserving data mining is a new research trend in privacy data for data mining and statistical database. Association analysis is a powerful tool for discovering relationships which are hidden in the large database. Association rule hiding algorithms get strong and efficient performance for protecting confidential and crucial data. Data modification and rule hiding are one of the most important approaches for secure data. Privacy-preserving data mining (PPDM) considers the problem of maintaining the privacy of data and knowledge in data mining.

In data mining, association rules are useful for analyzing and predicting customer behavior. They play an important part in shopping basket data analysis, product clustering, and catalog design and store layout. Programmers use association rules to build programs capable of machine learning. Machine learning is a type of artificial intelligence (AI) that seeks to build programs with the ability to become more efficient without being explicitly programmed.

The problem of association rule mining is defined as:

Let I = $\{i_1,i_2,i_3,\ldots\ldots.,i_n\}$ be a set of n binary attributes called items.

Let D = $\{t_1,t_2,t_3,\ldots\ldots.t_m\}$ be a set of transactions called the database.

Each transaction in D has a unique transaction ID and contains a subset of items in me.

A rule is defined as an implication of the form:

$X \Rightarrow Y$, where $X,Y \in I$

a rule is defined only between a set and a single item, $X \Rightarrow i_j, i_j \in I$.

Every rule is composed of two different sets of items, also known as item sets. X and Y, where X is called antecedent or left-hand-side (LHS) and Y is consequent or right-hand-side (RHS).

The support of a rule, X → Y, is the percentage of transactions in T that contains, and can be seen as an estimate of the probability; the rule support thus determines how frequent the rule is applicable in the transaction set T. Let n be the number of transactions in T. The support of rule X →Y is computed as follows:

$$\text{support} = \frac{|X \cup Y| * \text{count}}{N} \quad \text{---------------} \ (1)$$

Since N is constant (as it is the number of transactions in the given database). Support is a useful measure because if it is too low, the rule may just occur due to chance. Furthermore, in a business environment, a rule covering too few cases (or transactions) may not be useful because it does not make business sense to act on such a rule (not profitable).

Confidence: Confidence (strength or evidence) derives from a subset of the transaction in which two entities (or activities) are related. The confidence of a rule, X → Y, is the percentage of transactions in T that contain X also contain Y.

$$\text{confidence} = \frac{|X \cup Y| * \text{count}}{|X|} \quad \text{-----------}(2)$$

Confidence thus determines the predictability of the rule. If the confidence of a rule is too low, one cannot reliably infer or predict Y from X. A rule with low predictability is of limited use.Association rules are usually required to satisfy a user-specified minimum support and a user-specified minimum confidence at the same time. Association rule generation is usually split up into two separate steps:

1) First, minimum support is applied to find all frequent item sets in a database.

2) Second, these frequent item sets and the minimum confidence constraint are used to form rules.

2. PRIVACY PRESERVATION IN DATAMINING

Data Mining is a technique that extracts hidden predictive information from large volumes of data bases. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. With the amount of data doubling each year, more data is gathered and data mining is becoming an increasingly important tool to transform this data into information. Huge volumes of detailed personal data are regularly collected and analyzed by applications. Such Data include shopping habits, criminal records, medical history, and credit records, among others. Analyzing such data opens new threats to privacy .Privacy-preserving data mining (PPDM) is one of the important area of data mining that aims to provide security for secret information from unsolicited or unsanctioned disclosure. Data mining techniques analyzes and predicts useful information. In Fig1.1 the concept of privacy preserving data mining is primarily concerned with protecting secret data against unsolicited access. It is important because now a day's threat to privacy is becoming real since data mining techniques are able to predict high sensitive knowledge from huge volumes of data.

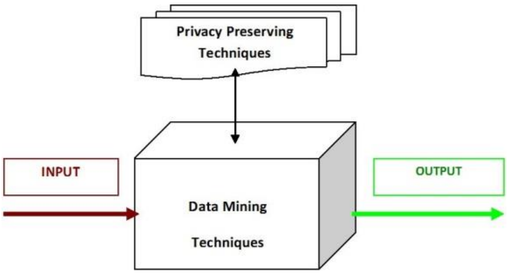


Fig 1.1: Privacy Preservation

Preservation of privacy in data mining has emerged as an absolute prerequisite for exchanging confidential information in terms of data analysis, validation, and publishing. Ever-escalating internet phishing posed a severe threat to the widespread propagation of sensitive information over the web. Conversely, the dubious feelings and contentions mediated unwillingness of various information providers towards the reliability protection of data from disclosure often results in utter rejection in data sharing or incorrect information sharing [9].

Privacy preserving data mining is a new research trend in privacy data for data mining and statistical database. Association analysis is a powerful tool for discovering relationships which are hidden in the large database. Association rules hiding algorithms get strong and efficient performance for protecting confidential and crucial data. Data modification and rule hiding are one of the most important approaches for secure data. Privacy-preserving data mining (PPDM) considers the problem of maintaining the privacy of data and knowledge in data mining.

The goal of privacy preserving data mining is to develop algorithms for modifying the original data in some way so that the private data and private knowledge remain private even after the mining process. Rule mining is used for finding frequent patterns, associations, correlations, or structures among sets of items or objects in transaction databases, relational databases, and other information repositories. It involves analyzing and presenting strong rules discovered in databases using different interest measures.

3. ARCHITECTURE

The architecture of the framework (fig 2.1) can be explained from the dataset collection from the specified data mining repository and the algorithm for association rule hiding is applied to the dataset and the frequent itemsets for combinations and sensitive rules are generated.

The hiding algorithm (FHSAR) is applied and those sensitive rules are hidden and modified dataset is obtained.

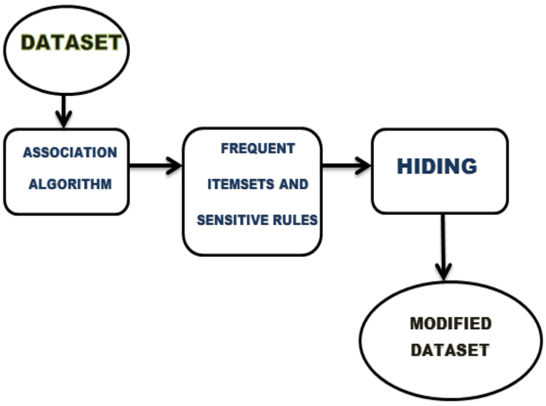


Fig 2.1: Design of the Framework

4. PROPOSED METHOD

The proposed framework (fig 3.1) is to generate high sensitive rules for the sensitive profile attribute/s based on minimum support and confidence using FHSAR hiding algorithm on frequent patterns considering the attributes, providing the minimum support 80% and confidence 80% and generate the frequent item sets and generate sensitive rules for the dataset and then hide the sensitive rules generated for the attributes in each transaction in the dataset and modify the original dataset.

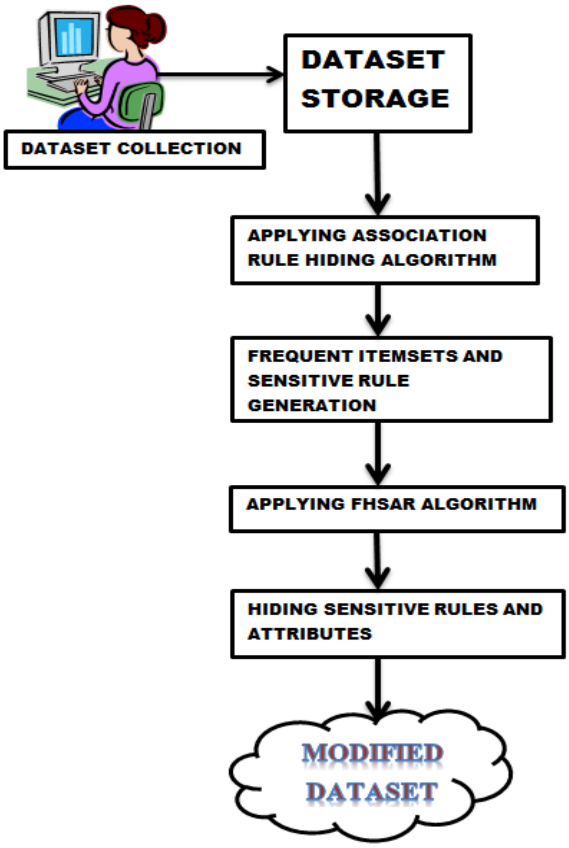

DATASET STORAGE
DATASET COLLECTION
APPLYING ASSOCIATION RULE HIDING ALGORITHM
FREQUENT ITEMSETS AND SENSITIVE RULE GENERATION
APPLYING FHSAR ALGORITHM
HIDING SENSITIVE RULES AND ATTRIBUTES
MODIFIED DATASET

Fig 3.1: Proposed Framework

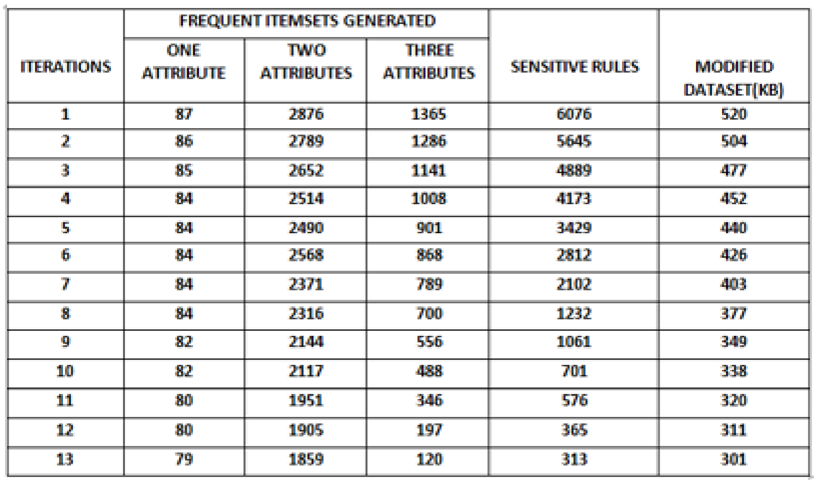
```
FHSAR( );
Input: D, SAR, min_support, min_confidence;
Output: D', where all SAR will be hidden.
Stage 1
For each transaction ti in the database D Do
  { Foreach sensitive rule SARj∈ SAR Do
    { IfSARj supported by tiThen
      { ||SARj|| = ||SARj|| + 1;
        ||L(SARj)|| = ||L(SARj)|| + 1;
      }
      Else IF L(SARj) suppoted by tiThen
        ||L(SARj)|| = ||L(SARj)|| + 1;
    }
    If exist any SARj supported by ti
    { MICi = Item_Selection ( );
      wi= MICi / 2 ( |ti | - 1);
      Store the ti's ID and the wi in PWT;
    }
  }
Stage 2
While SAR is not empty (≠∅) do
  { SelecttID from PWT with maximal weight w;
    tID.k = Item_Selection ( );
    IF Checking_and_RemovingItem( ) == True Then
    { ModifywID of the tID and insert the ID into the PWT
      in the maintained order;
      For each SARj that tID.k∈SARjDo
      { IFSARj⊆tIDThen
          ||SARj|| = ||SARj|| – 1;
        IF (L(SARj)⊆tID) and ((tID.k)∈L(SARj) ) Then
          ||L(SARj)|| = ||L(SARj)|| – 1;
        IF (support(SARj) <min_support) or
           (confidence(SARj) <min_confidence) Then
          Remove SARj from SAR;
      }
    }
  }
```

The pseudo code for FHSAR algorithm used to hide the sensitive rules that get generated using the association rule hiding algorithm.

5. PERFORMANCE EVALUATION

The dataset connect is obtained from SPMF an open source data mining library prepared based on real-life data [10]. The Frequent item sets for the dataset with minimum support and confidence 80% is generated and sensitive rules are obtained and stored in the files. The file storing the transaction number and the attributes that have to hide from each transaction in the original dataset. The modified dataset (504KB) after hiding the attributes from each transaction in the original dataset (540KB) is obtained. For each iteration, modified dataset is considered and the frequent item sets, sensitive rules and the attribute to be hidden for each transaction is processed after 13th iteration the rules generated reduced to 313 rules as shown in Table 4.1.

Table 4.1: Iteration Table



ITERATIONS	FREQUENT ITEMSETS GENERATED			SENSITIVE RULES	MODIFIED DATASET(KB)
	ONE ATTRIBUTE	TWO ATTRIBUTES	THREE ATTRIBUTES		
1	87	2876	1365	6076	520
2	86	2789	1286	5645	504
3	85	2652	1141	4889	477
4	84	2514	1008	4173	452
5	84	2490	901	3429	440
6	84	2568	868	2812	426
7	84	2371	789	2102	403
8	84	2316	700	1232	377
9	82	2144	556	1061	349
10	82	2117	488	701	338
11	80	1951	346	576	320
12	80	1905	197	365	311
13	79	1859	120	313	301

6. CONCLUSION

This Paper focuses on a framework for hiding the sensitive data attributes of the dataset connect from SPMF an open source data mining library.The frequent item sets and the sensitive rules are generated using the association mining rule called apriori. By using Fastest hiding sensitive association rule (FHSAR) hiding algorithm these sensitive rules are hided and the dataset is modified. In the future, the efficiency of the algorithm can be tested with other algorithms.

7. REFERENCES

[1] Yi-Hung Wu., Senior Member, IEEE Computer Society, “Hiding Sensitive Association Rules with Limited Side Effects”.IEEE transactions on knowledge and data engineering, January 2007, vol. 19, no. 1
[2] Chih-Chia Weng, “A Novel Algorithm for Completely Hiding Sensitive Association Rules” ,Eighth International Conference on Intelligent Systems Design and Applications 2008 IEEE, 978-0-7695-3382-7/08
[3] Wang, S.-L.,Kuan-Wei Huang, Tien-Chin Wang et al. “Maintenance of discovered informative rule sets: incremental deletion. in Systems”, Man and Cybernetics, 2005 IEEE
[4] Peng Cheng, John F. Roddick, Shu-Chuan Chu•, Chun-Wei Lin, “Privacy preservation through a greedy, distortion-based rule-hiding method”, Springer Science Business Media New York 2015
[5] Murat Kantarcioglu, “Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data”, IEEE transactions on knowledge and data engineering, September 2004 vol. 16, no. 9
[6] Amiri , A., “Dare to share: Protecting sensitive knowledge with data sanitization. Decision Support Systems”, 2007. 43(1): p. 181-191
[7] M.BalaGanesh , Mrs.V.Sathya, S.Vinoth Kumar, “A Secure Framework for Protection of Social Networks from Information Stealing Attacks “,IJIRCCE Vol. 2, Issue 1, January 2014
[8] R. Agrawal and R. Srikant, “Fast Algorithms for Mining Association Rules,” Proc. 20th International Conference. Very Large Data Bases, pp. 487-499, 94,http://www.vldb.org/dblp/db/ conf/vldb/vldb94-487.html
[9] Jiuyong Li, Hong Shen, Rodney Topor, “Mining the Smallest Association Rule Set for Predictions”, Proceedings of the 2001 IEEE International Conference on Data Mining, pp.361-368, 2001.
[10] https://springerplus.springeropen.com/articles/10.1186/s40064-015-1481-x
[11] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, “PrivacyPreserving Mining of Association Rules,” Information Systems, vol. 29, pp. 343-364, 2004.
[12] http://www.philippe-fournier-viger.com/spmf/datasets/connect.txt
[13] http://www.philippe-fournier-viger.com/spmf/SensitiveAssociationRules.php
[14] http://www.borgelt.net/apriori.html, May 2007
[15] http://www.msisac.org/