Volume 2, Issue 10, October 2012 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

An Extended Method for Privacy Preserving Association Rule Mining

Shikha Sharma

PG Student,

Shri Vaishnav Institute of Technology and Science,

Indore (MP), India

Abstract: ***The protection of information from illegal access has been a long term goal for businesses and government organizations. Recent progresses in the field of the data mining algorithms have expands the risk while releasing data outside. The key problem is not still inspected; there is the need to balance privacy of disclosed data with the appropriate need of data users. Every disclosure method affects the data and modifies true value and relationship. In this thesis we remove this drawback. In other techniques non-sensitive rules hidden (falsely) as a side effect and artificial rules falsely generated in other rule hiding techniques. In this paper we present a new approach that necessarily changes few transactions in the transaction database by decreasing support or confidence of sensitive rules without any side effect.***

Keywords: Privacy Preserving Data Mining, Association Rule Mining, Sensitive Rule Hiding.

I. Introduction

Data mining is the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable pattern in data. Data mining represents the integration of several fields, including machine learning, database systems, data visualization, statistics and information theory. Several major data mining techniques have been developed and used in data mining projects recently including association, classification, clustering, prediction and sequential patterns [1]**.**

Knowledge discovery in databases is a complex process, which covers many interrelated steps. Key steps in the knowledge discovery process are-

DATA SELECTION- The data needed for the data mining process may be obtained from many different and heterogeneous data sources. This first step obtains the data from various databases, files and non-electronic sources.

DATA PREPROCESSING- The data to be used by the process may have incorrect or missing data. There may be anomalous data from multiple sources involving different data types and metrics. There may be many different activities performed at this time. Erroneous data may be corrected or removed, whereas missing data must be supplied or predicted.

DATA TRANSFORMATION- Data from different sources must be converted into common format for processing. Some data may be encoded or transformed into more usable formats. Data reduction may be used to reduce the number of possible data values being considered.

DATA MINING- This step applies algorithms to the transformed data to generate the desired results.

DATA PATTERN EVALUATION/INTERPRETATION-How the data mining results are presented to the users is extremely important because the usefulness of the results is dependent on it. Various and GUI strategies are used at this last step.
Knowledge Discovery in Databases (KDD) is an automated extraction of novel, understandable and potentially useful patterns implicitly stored in large databases, data warehouse and other massive information repositories. KDD is a multi-disciplinary field drawing work from areas including database technology, artificial intelligence, machine learning, neural networks, statistics, pattern recognition, information retrieval, high performance computing and data visualization [1].

Privacy preserving data mining [2, 3] is a new investigation in data mining and statistical databases [4]. In PPDM data mining algorithms are analyzed for side effects obtain in data privacy. Two fold consideration in privacy preserving data mining. First is sensitive raw data that are kept secure from unauthorized access like identifiers, names ,addresses should be modified from original database in order for receiver of data not to be able to compromise another person's privacy. Second is sensitive knowledge is excluded that can be mined from a database by using data mining algorithms as such type of knowledge compromises data privacy[4].

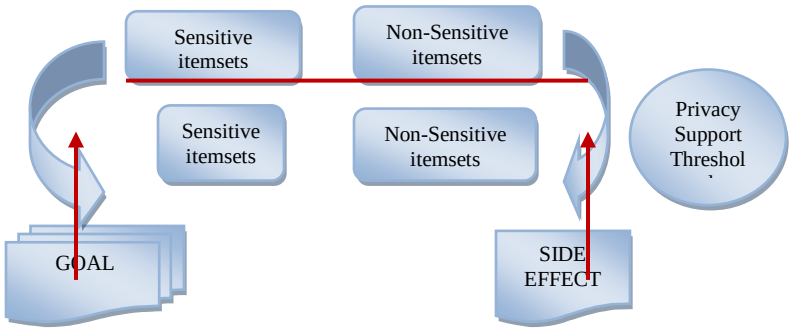

Figure 1.1: The goal and side effects of PPDM for association rule mining

GOAL- The goal is to reduce support of sensitive itemsets below privacy support threshold.
SIDE EFFECT- The lower we reduce support of sensitive itemsets, the more we reduce support of non-sensitive itemsets, which might go below privacy support threshold.

II. Related Work

There is a description of the literature relevant to our field and topic, for the purpose of survey to understand the domain of problem and possible approaches to solve the problem and improvements suggested by different authors so that it is possible to design the new algorithm which may solve the problems with some improvements. In this section, the summary of different algorithm, problems associated with them and the different approaches used by different authors to solve those problems are discussed. As suggested by V.S. Verykios, A.K. Elmagarmid, B. Elisa, Y. Saygin, and D. Elena [7], there is large depository of data that contains sensitive data that must be protected from illegitimate access. As we know that protection of data from illegal access is a long term goal for database security research community, governmental organizations and business associations. As advancement in data mining arises with that risk of releasing data to outside party also increases. Every disclosure method affects data in some way and modifies to true value and relationship. In this paper, they investigate confidentiality of a broad category of association rules. In this they presented algorithms and strategies for hiding a group of association rules is characterized as sensitive. A rule is sensitive if its disclosure risk is above certain privacy threshold. They perform an evaluation study of the hiding algorithms in order to analyze their time complexity and the impact that they have in the original database. The security impact of data mining is analyzed 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 adopts an approach in that he studied the correlation between the amount of released data and the significance of the patterns that were discovered.

They also show how to determine the sample size in such a way that data mining tools cannot obtain reliable results. Clifton and Marks is 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 is independent from any specific data mining technique; other researchers 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 apriori. The authors in 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. In this research they presented two fundamental approaches in order to protect sensitive rules from disclosure. The first approach prevents rules from being generated by hiding the frequent sets from which they were derived. The second approach reduces the importance of the rules by setting their confidence below a user-specified threshold. In this research they presented two fundamental approaches in order to protect sensitive rules from disclosure.

As suggested by Kazem Taghva, Pavankumar Bondugula, Darshana Gala [6]. An association rule expresses the dependency of a one set of attributes on another attributes. In the identification of association rules, an Apriori algorithm is one of the known techniques that is used. In this research, this technique is used for privacy data identification and extraction from printed documents. In this research they point the problem of discovering association rules for various privacy types from printed documents. An association rule expresses the dependence of a set of attribute-value pairs and upon another set of items (itemset). The mining of association rules is performed in two stages: The frequent sets of items from the data discovery and association rules generation from the frequent item sets. Searching of these frequent itemsets is in general a

combinatorial expensive task. Association rule mining has a broad range of applicability. It was first introduced to find the association between items in supermarket transactions for promotion of sales, arrangement of associated items accordingly, to increase profits etc

III. Proposed Strategy

To hide any specified association rule X → Y this algorithm works on the basis of confidence (X → Y) and support (X → Y).To hide the rule X → Y (containing sensitive element X on LHS),this algorithm increases the special variable of the rule X → Y until confidence (X → Y) goes below a minimum specified threshold confidence (MCT).As the confidence (X → Y) goes below MCT(minimum specified confidence threshold),rule X → Y is hidden i.e. it will not be discovered through data mining algorithm.

***INPUT*:**

1: A source database D.
2: A set of strong rules
3: MST (Minimum Support Threshold).
4: MCT (Minimum Confidence Threshold).
5: A set of sensitive items X.
6: A set of special variable for all rules (which are initially set to zero).
7: New modified terms confidence (X → Y) and support (X → Y).

***OUTPUT*:**

A transformed database of strong rules with modified support and confidence where rules containing X on LHS or RHS will be hidden.

PROCEDURE:

Initially set the special variables of all the rules equal to 0.
Il check for all sensitive elements.
for each x in X where x belongs to X
{
// Now check all the rules containing sensitive element x.
For each rule R which contain x on LHS or RHS
{
II Check whether Modified confidence of the rule
Il goes below MCT or not.
Special variable = 0

Compute support and confidence using:

$$\text{Confidence}(X \rightarrow Y) = \frac{I_{X \cup Y}}{I_X + SV(X \rightarrow Y)}$$

$$\text{Support}(X \rightarrow Y) = \frac{I_{X \cup Y}}{N + SV(X \rightarrow Y)}$$

While (confidence (R > = MCT)
& (support(R > = MST)
II increase the special variable of
II rule R by 1
{
Special variable (R) = special variable + 1
}
}
}
End of procedure

IV. Performance Evaluation

Association rule mining over woman's clothing store [3] is considered a basic knowledge discovery activity. For discovering correlations among items, Association rule mining provides a useful mechanism belonging to customer transactions in a woman's clothing store database. Let D be the database of transactions and I= {I1, ..., In} be the set of items. A transaction T includes one or more items in I .An association rule has the form A→ B, where A and B are non-empty sets of items (i.e. A and B are subsets of I) such that A ∩ B = Null. A set of items is called an itemset, while A 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 A → B is the ratio of the number of transactions that contain A or B to the number of transactions that contain A.

4.1 SOLUTION BY PROPOSED METHOD

We take an example of woman's clothing store in which we are having four items {Jeans, T-shirt, Skirt, Shoes} and five transactions [4]. We assume minimum support threshold (MST) of 60% and minimum confidence threshold (MCT) of 70%.

TABLE 6.1- Transaction Data Table I

TID	ITEMS
T1	JEANS,TSHIRT,SHOES
T2	TSHIRT
T3	JEANS,SKIRT,SHOES
T4	JEANS,TSHIRT
T5	JEANS,TSHIRT,SHOES

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

JEANS->TSHIRT (60%, 75%)
TSHIRT->JEANS (60%, 75%)
JEANS->SHOES (60%, 75%)
SHOES->JEANS (60%, 100%)
Now there is a need to hide TSHIRT and SHOES as it is sensitive.

TABLE 6.2- Initial Association Rule Constraints Data Table

	SUPPORT	CONFIDENCE	SV
JEANS->TSHIRT	60%	75%	0
TSHIRT->JEANS	60%	75%	0
JEANS->SHOES	60%	75%	0
SHOES->JEANS	60%	100%	0

The above table contains four rules and the graph is shown below-

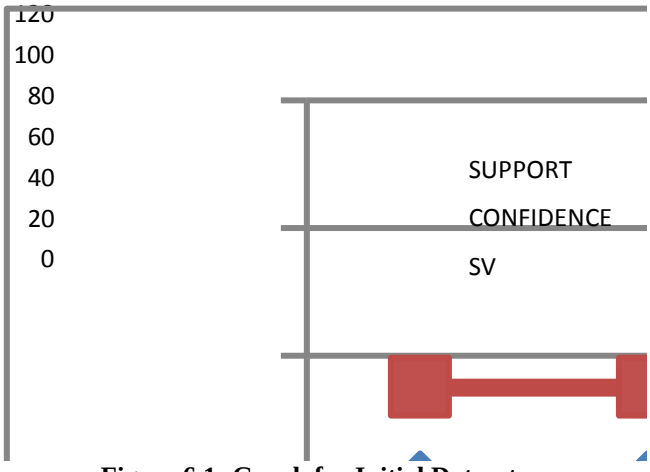

Figure 6.1- Graph for Initial Dataset

4.1.1 Approach to hide TSHIRT

TABLE 6.3- Transaction Data Table II

TID	ITEMS
T1	JEANS,TSHIRT,SHOES
T2	TSHIRT
T3	JEANS,SKIRT,SHOES
T4	JEANS,TSHIRT
T5	JEANS,TSHIRT,SHOES

TABLE 6.4- Data Table hiding TSHIRT

	SUPPORT	CONFIDENCE	SV
JEANS->TSHIRT	60%	75%	0
TSHIRT->JEANS	**48%**	**58%**	**1**
JEANS->SHOES	60%	75%	0
SHOES->JEANS	60%	100%	0

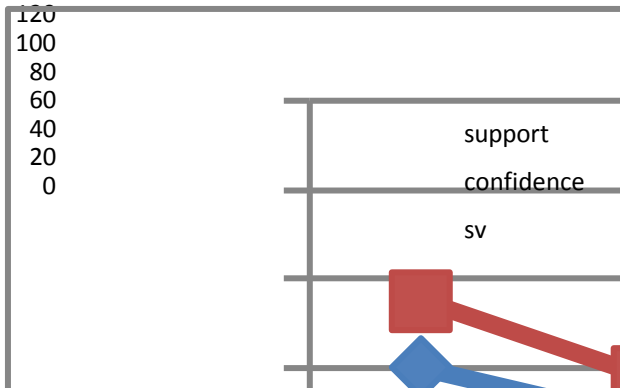

Figure 6.2- Graph for hiding rules I

The above graph shows the values of support and confidence that reduces hiding sensitive element TSHIRT.

4.1.2 Approach to hide SHOES

TABLE 6.5- Transaction Data Table III

TID	ITEMS
T1	JEANS,TSHIRT,SHOES
T2	TSHIRT
T3	JEANS,SKIRT,SHOES
T4	JEANS,TSHIRT
T5	JEANS,TSHIRT,SHOES

The above table contains five transactions. In this approach we hide shoes(using LHS) using hiding algorithm, so for hiding shoes hidden counter runs two times. So we get the values of support and confidence below minimum support threshold and minimum confidence threshold. So by our approach the rule for jelly is hidden as shoes is sensitive element.

TABLE 6.6- Data Table hiding SHOES

	SUPPORT	CONFIDENCE	SV
JEANS->TSHIRT	60%	75%	0
TSHIRT->JEANS	**48%**	**58%**	**1**
JEANS->SHOES	60%	75%	0
SHOES->JEANS	**42%**	**58%**	**2**

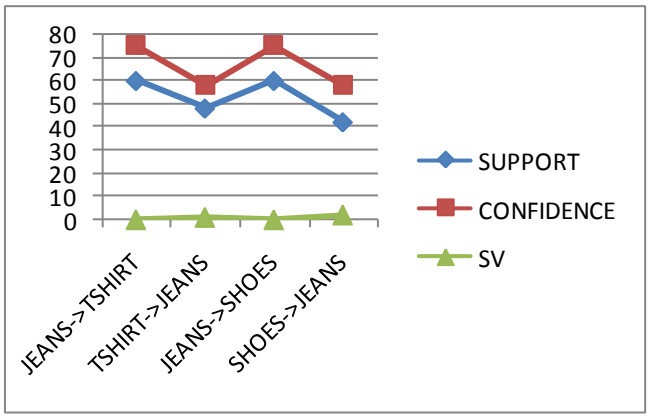

Figure 6.3- Graph for hiding rules II

So it is clear that this approach is hiding all the given sensitive rules successfully without any side effect.

V. Conclusion

From experimental results we see that our approach is better in the way that it hides any rule which cannot be hidden by some of the previous works. We see in the example that proposed method is hiding the given association rules TSHIRT->JEANS (with sensitive items on the left hand side of the rule) without any side effect. The aim of this research release non-sensitive item sets while keeping sensitive items private.

REFERENCES

[1] M.H.DUNHAM, DATA MINING Introductory and Advanced Topics(01-sep-2006) , by Pearson Education.

[2] M. Kantarcioglu and C. Clifton. *Privacy-preserving distributed mining of association rules on horizontally partitioned data*. In IEEE Transactions on Knowledge and Data Engineering Journal, volume 16(9), pages 1026–1037, Piscataway, NJ, USA, September 2004. IEEE Educational Activities Department.

[3] 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.

[4] R..Agrawal, T.. mielinski, and A.Swami.*Miningassociation rules between sets of items in large databases*. In Proceedings of the ACM SIGMOD Conference on Management of Data, pages 207–216, New York, NY, USA, May 1993. ACM Press.

[5] V. S. Verykios, A. K. Elmagarmid, B. Elisa, Y. Saygin, and D. Elena. *Association rule hiding*. In IEEE Transactions on Knowledge and Data Engineering, volume 16(4), pages 434–447, Los Alamitos, CA, USA, April 2004. IEEE Computer Society.

[6] Kazem Taghva, Pavankumar Bondugula,DarshanaGala .*Association Rule Discovery of Privacy Information.* Information Science Research Institute, University of Nevada, Las Vegas NV 89154-4021

[7] Vassilios S. Verykios, Ahmed K. Elmagarmid, Bertino Elisa, Yucel Saygin, and Dasseni Elena, Association Rule Hiding, IEEE Transactions on Knowledge and Data Engineering (2003), Accepted

[8] Mike J. Atallah, Elisa Bertino, Ahmed K. Elmagarmid, Mohamed Ibrahim, and Vassilios S. Verykios, Disclosure Limitation of Sensitive Rules, In Proceedings of the IEEE Knolwedge and Data Engineering Workshop (1999), 45–52.

[9] Elena Dasseni, Vassilios S. Verykios, Ahmed K. Elmagarmid, and Elisa Bertino, Hiding Association Rules by using Confidence and Support, In Proceedings of the 4th Information Hiding Workshop (2001), 369–383

[10] Yucel Saygin, Vassilios S. Verykios, and Ahmed K. Elmagarmid, Privacy preserving association rule mining, In Proceedings of the 12th International Workshop on Research Issues in Data Engineering (2002), 151–158

[11] Yehuda Lindell and Benny Pinkas, Privacy preserving data mining, In Advances in Cryptology - CRYPTO 2000 (2000), 36–54.

