Using Data Mining Confidence and Support for Privacy Preserving Secure database

Prof. Dr. Alaa H. AL-Hamami
Alaa_hamami@yahoo.com, Amman Arab University for Graduate Studies, Zip Code: 11953, P.O.B. 2234, Amman, Jordan, 2008.
Dr. Mohammad A. AL-Hamami
M_ah_1@yahoo.com,
Delmon ,Bahrain, 2008
Dr. Soukaena H. Hashem
soukaena_hassan@yahoo.com,
University of technology, Iraq, 2008

ABSTRACT

Data mining is a process that uses a variety of data analysis tools to discover patterns and relationships that can be hidden among vast amount of data. The results patterns of Data Mining (DM) such as association rules, classes, clusters, etc, will be readily available for working team. So the mining will penetrate the privacy of sensitive data and makes the stolen of the knowledge resulted much more easily.

This research introduces suggestion to secure the results of the data Mining. For privacy preserving secure databases we aim to hide the general secure and sensitive rules from appearing as a result of applying Association Rules techniques. This could be done by making the confidence of secure rules equal to zero by modifying the supports of critical and sensitive items in these rules.

Keywords:
Data mining, Association Rules, Sensitive rules, Support, and Confidence.

1. Introduction.

Data mining derives its name from the similarities between searching for valuable information in a large database and mining rocks for a vein of valuable ore. The more general terms such as Knowledge Discovery in Databases (KDD) describe a more complete process. Data mining is being put into use and studied for databases, including relational databases, object-relational databases and object oriented databases, data warehouses, transactional databases, unstructured and semi structured repositories such as the World Wide Web, advanced databases such as spatial databases, multimedia databases, time-series databases and textual databases, and even flat files [1].

The data mining functionalities and the variety of knowledge they discover are briefly presented in the following list: **Characterization**: Data characterization is a summarization of general features of objects in a target class, and produces what is called *characteristic rules*. **Discrimination**: Data discrimination produces what are called *discriminate rules* and is basically the comparison of the general features of objects between two classes referred to as the *target class* and the *contrasting class*. **Association analysis**: Association analysis is the discovery of what are commonly called *association rules*. It studies the frequency of items occurring together in transactional databases, and based on a threshold called *support*, identifies the frequent item sets. Another threshold, *confidence*, which is the conditional probability than an item appears in a transaction when another item appears, is used to pinpoint association rules. **Classification**: Classification analysis is the organization of data in given classes. Also known as *supervised classification*, the classification uses given class labels to order the objects in the data collection. Classification approaches normally use a *training set* where all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model. The model is used to classify new objects. **Clustering**: Similar to classification, clustering is the organization of data in classes. However, unlike classification, in clustering, class labels are unknown and it is up to the clustering algorithm to discover acceptable classes [2, 3].

2. Association Rules Algorithm.

The efficient discovery of such rules has been a major focus in the data mining research community. Many algorithms and approaches have been proposed to deal with the discovery of different types of association rules discovered from a variety of databases. However, typically, the databases relied upon are alphanumerical and often transaction-based. The problem of discovering association rules is to find relationships between the existence of an object (or characteristic) and the existence of other objects (or characteristics) in a large repetitive collection.

The problem is stated as follows, see table 1, Let $I = \{i1, i2, ...im\}$ be a set of literals, called items. Let D be a set of transactions, where each transaction T is a set of items such that $T \subseteq I$. A unique identifier *TID* is given to each transaction. A transaction T is said to contain X, a set of items in I, if $X \subseteq T$. An *association rule* is an implication of the form "$X \Rightarrow Y$", where $X \subseteq I$, $Y \subseteq I$, and $X \cap Y = \emptyset$. The rule $X \Rightarrow Y$ has a *support s* in the transaction set D is s% of the transactions in D contain $X \cup Y$. In other words, the support of the rule is the probability that X and Y hold together among all the possible presented cases. It is said that the rule $X \Rightarrow Y$ holds in the transaction set D with *confidence c* if c% of transactions in D that contain X also contain Y. In other words, the confidence of the rule is the conditional probability that the consequent Y is true under the condition of the antecedent X. The problem of discovering all association rules from a set of transactions D consists of generating the rules that have a *support* and *confidence* greater that given thresholds. These rules are called *strong rules [4]*.

Such a repetitive collection can be a set of transactions for example, also known as the market basket. Typically, association rules are found from sets of transactions, each transaction being a different assortment of items, like in a shopping store ({milk, bread, etc}). Association rules would give the probability that some items appear with others based on the processed transactions, for example milk →bread [50%], meaning that there is a probability 0.5 that bread is bought when milk is bought. Essentially, the problem consists of finding items that frequently appear together, known as frequent or large item-sets.

Table 1: A model of a simple transaction database

TID	Items
001	A C D
002	B C E
003	A B C E
004	B E

3. The Proposed Solution.

The main objective of the proposed system is preserving the privacy of data mining. There are many researches discuss this application, that by developing algorithms for modifying, encrypting and distributing the original data in the database to be mined. The privacy of data (original data in database that will be mined) and the privacy of knowledge (the association rules extracted from mined database) will be ensured even after the mining process has taken place. The problem that arises when confidential information can be derived from released data by unauthorized users can be solved. As an example of database that has sensitive data, the Employers databases will be introduced as in Figure (1).

TID	Name	Age	Address	Marriage state	No.of child	Owns	Politic orientation	Income $	Account
1	Suzan	30	Baghdad	Marriage	3	3 house , car	has	1000	4555
2	Ahmad	35	Baghdad	Devisor	6	House, car	hasn't	67	54432987
.									
.									

Figure 1: Employers database

After the encoding of the Employers database, it will be as follows:

- Both name and account will be cancelled in mining process and will be known by no. of TID.
- For ages more than 40 presented by A else A will not appear. For address Baghdad presented by B else B will not appear. Marriage presented by C else C will not appear. No. of child more than three presented by D else D will not appear. Owns more than car and one house presented by E else E not appear. Politic orientation if has presented by F else F not appear. Finally Income more 500$ presented by G else G will not appear.

The proposed solution is to hide sensitive rules by reducing the confidence of these rules. Suppose the following:

- The rule is A → B.
- Confidence (A → B) = support (AB) / support (A).
- Decreasing confidence of the rule by increasing the support of A in transactions without supporting B. Decreasing the support of B in transactions that supporting both A and B.
- Then surly the new Confidence of the rule (A → B) = zero.

4. Implementation.

Now to implement the suggestion, Figure 2 will present the form which introduces the textbox to put it in the name of the notepad file which contains the data of encoded Employer's database. Then there is a command called mine which mines the data. If the last has been clicked then the apriori association rule algorithm will be executed. The resulted association's rules will be saved in notepad file and its name will be displayed in other form as in Figure 3. This figure has a command called display association rules. When you clicked on this command , the notepad file which save the rules will appear as in Figure 4.

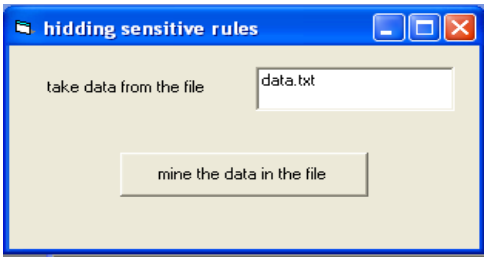

Figure 2: Window for implementing association rule mining

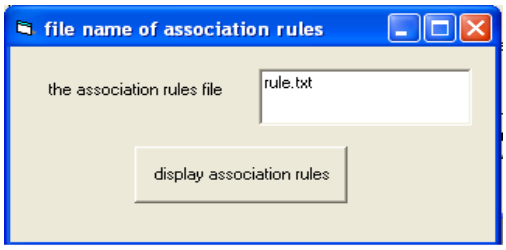

Figure 3: Window presents name of the association rules file

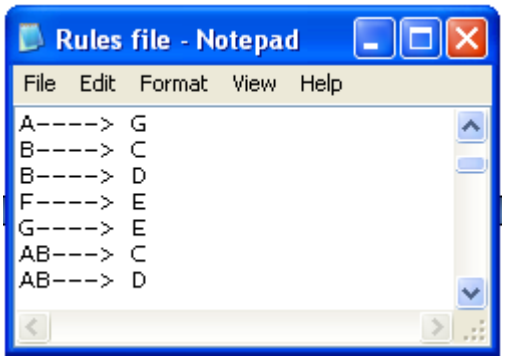

Figure 4: Window displays association rules

Now the analysis stage will begin and from the analysis we will get some of much more sensitive rules which penetrate the privacy of the employers. As an example of these rules is:

F--→ E (100%)
G-→ E (100%)
ABFG --→ E (100%)

- This rules present that, the employer whose age more than 30, live in Baghdad, his income more than 500$ and has political orientation surely with 100% confidence he has car and more than one house.
- These rules very sensitive since it predicts it is owns, since the state of income and political state are previously known.

To solve this problem, the following suggestion will be executed:

Since these analyzed association rules will be introduced to all the official employers to make their prediction into supporting the advance of employers, then we must hide these sensitive rules without any effects on other rules.

Now rebuild the employer database by the steps mentioned in the section number 3 in the suggestion which involve the rearrangement of attributes supports. After these modifications we see that:

F--→ E (0%)
G-→ E (0%)
ABFG --→ E (0%)

The very sensitive rules disappeared without effect on other normal rules, since the hidden rules are a general rules (related with more employers) but sensitive to them. See figure 5.

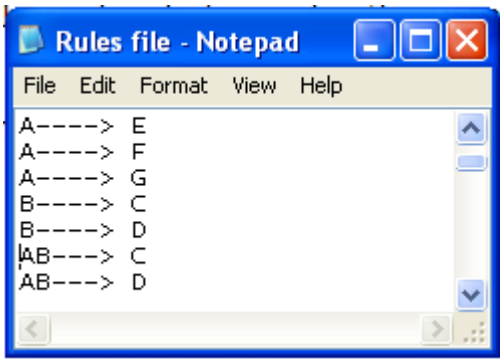

Figure 5: Window displays association rules after hiding sensitive rules

4. Conclusions.

From the proposed solution we conclude the following:

1. The suggestion which implies hiding rules those very sensitive to general employers by decrease the confidence of them to zero, play a big role in privacy preserving techniques.
2. By modifying database to hide sensitive rules we will guarantee space and time. Since the rule extraction computation done in limited time, also the space needed to store the rules will be decreased.
3. Really hiding sensitive rules will decrease the no. of resulted rules then the analysis stage optimized and reports will introduced for official employers will be abstracted and limited then the predictions built from the extracted patterns be more accurate.

References.

[1]. Arron Ceglar, John Roddick; "***Association Mining"***; ACM Computing Surveys, Vol. 38, No. 2, Article 5, pp. 1-42, July 2006.

[2]. Kantardzic M.; ***"DM concepts, models, methods and algorithms"***, jhon wiley &Sons, 2003.

[3]. Ye N., ***"The Handbook of Data Mining"***, Edited by Human Factors and Ergonomics, www.erlbaum.com, 2001.

[4]. Berry M. and Linoff G., ***"Data Mining Techniques"***, John Wiley, 1997.

[6]. Ala'a H. AL-Hamami, Mohammad Ala'a Al-Hamami and Soukaena Hassan Hasheem, ***" Applying data mining techniques in intrusion detection system on web and analysis of web usage"***, Asian Journal of Information Technology, 2006.

[7]. Ala'a H. AL-Hamami, and Soukaena Hassan Hasheem***," Privacy Preserving for Data Mining Applications"***, journal of technology, baghdad, Iraq, university of technology, 2008.

[8]. Mohammad A. Al- Hamami and Soukaena Hassan Hashem, ***" Applying Data Mining Techniques to Discover Methods that Used for Hiding Messages Inside Images "***, The IEEE First International Conference on Digital Information Management (ICDIM2006), Bangalore, India, 2006.