Privacy Preserving for Data Mining Applications

Prof. Dr. Ala'a H. AL-Hamami*   Dr. Soukaena Hassan Hashem**

Received on:16/5/2007
Accepted on:6/3/2008

Abstract

The results 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. The main objective of the proposed system is preserving the privacy of data mining, that will done by developing algorithms for modifying, encrypting and distributing the original data in the database to be mined. So we ensure the privacy of data (original data in database that will be mined) and the privacy of knowledge (the association rules extracted from mined database) 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.

Keyword:
Data mining, privacy, sensitive databases, protection techniques, Twofish encryption.

الخلاصة:

نتائج خوارزميات تعدين البيانات مثل القواعد المترابطة والعنقدة والتصنيفات كلها تكون متوفرة ومعلنة لفريق العمل الذي قام باستخراج هذه النتائج. وبهذا فأن عملية التعدين تسبب سهولة اختراق المعلومات الحساسة وتسهل عملية سرقة المعارف الناتجة منها. ان الهدف الرئيسي من هذا البحث هو حماية خصوصية البيانات المراد تعدينها من خلال بناء خوارزميات تحوير وتشفير وتوزيع البيانات الاصلية الموجودة ضمن قاعدة البيانات المراد تعدينها. وبهذا نضمن خصوصية البيانات (البيانات الاصلية في القاعدة المراد تعدينها) وخصوصية المعارف المستحصلة (خصوصية القواعد المترابطة الناتجة من قاعدة البيانات المعدنة) بعد انتهاء عملية التعدين. المشكلة تتنامى عندما تخترق موثوقية المعلومات من قبل الاشخاص غير المخولين بسبب كون المعلومات متوفرة ومعلنة بشكل صريح.

https://doi.org/10.30684/ etj.29.5.8
University of Technology-Iraq, Baghdad, Iraq/2412-0758

Introduction.

With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, if not necessary, to develop powerful means for analysis and perhaps interpretation of such data and for the extraction of interesting knowledge that could help in decision-making. In principle, data mining is not specific to one type of media or data. Data mining should be applicable to any kind of information repository. However, algorithms and approaches may differ when applied to different types of data. Data mining is being put into use and studies for databases, including relational databases, object-relational databases and object oriented databases, data warehouses, transactional databases. Also include 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. The kinds of patterns that can be discovered depend upon the data mining tasks employed. The data mining functionalities and the variety of knowledge they discover are briefly presented in the following list: [1-6].

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. The proposed system.

A proposed research area of privacy preserving for data mining is proposed. Recent advances in data collection, data dissemination and related technologies have inaugurated a new era of research where existing data mining algorithms should be reconsidered from a different point of view, ***this of privacy***. Because the Internet and other media, have reached to a point where threats against the privacy are very common on a daily basis and they deserve serious thinking.

Privacy for data mining is a research direction in data mining, where data mining algorithms are analyzed for the side-effects they incur in data privacy. The main consideration in privacy for data mining is two fields:

- Sensitive data like identifiers, names, addresses and the like, should be modified, encrypted or distributed in the original database, that for making the recipient of the data not to be able to compromise another person privacy.
- Sensitive knowledge which can be mined from a database by using data mining algorithms should also be excluded, because such knowledge can equally well compromise data privacy.

Fig. (1) shows a database for bank customers that base record contains many of sensitive data for those customers such as identifiers, names, age, address, and marriage state, no. of children, income, credit information, and account. Fig. (2) shows a database for web log in/out which, has all the information about the logging of user from and to the specified web, so this information must be secure and if be mined will give a good pattern for detect the intrusion and web usage. So, these two databases must be protected and must have complete privacy preservation.

2.1. The modifying and encryption algorithms.

Modifying

Data modification is used in order to modify the original values of a selected attribute of database that needs to be released to the public and in this way to ensure high privacy protection. It is important that a data modification technique should be in concert with the privacy policy adopted by the administrators. Here three types of modification are proposed:

- Replacement, which is accomplished by replacing of an attribute value by a new value (i.e., changing a 1-value to a 0-value, or adding noise).
- Concatenation, which combines several values into a coarser category.
- Interchanging, which changes values of individual record with other one, such as replace the place of record 4 instead of record 10 and so on.

Encryption

Apply encryption algorithm for all values of selected attribute. This propsed research suggest encrypting the values by using symmetric encryption algorithm since it deal with static environment which has one parity, the administrators of the database, (here we will select twofish encryption algorithm [7]).

Twofish is a 128-bit block cipher that accepts a variable-length key up to 256 bits. The cipher is a 16-round Feistel network with a bijective *F* function made up of four key-dependent 8-by-8-bit S-boxes, a mixed 4-by-4 maximum distance separable matrix over GF(28), a pseudo-Hadamard transform, bitwise rotations, and a carefully designed key schedule.. Twofish can be implemented in hardware in 14000 gates.

Fig.(3) shows the databases after applying the three proposed modification techniques and the encryption algorithm on all the attributes of the bank database. Now will explain what happen for changes in the protected databases in the following steps:

1. The protection of the first attribute (the identifier) will be done by the proposed replacement by changing the digits of the identifier and that by adding random digit. at the first of it and two digits at the end of it, the changing schema as here (0 to 3, 1 to 2, 2 to 4, 3 to 6, 4 to 8, 5 to 7, 6 to 0, 7 to 1, 8 to 9, 9 to 5). For example the first identifier 500 (in figure 1) will be 733 by changing and finally after adding noise will be 473368.
2. The protection of the second, fourth, seventh, eighth and ninth attribute (name, address, income, credit information and account) will be done by encrypting the values by

using twofish encryption algorithm; see the original values in figure (1) and the encrypted values in figure (3). The values encrypted by implementing the twofish program, see fig.(4).

3. The protection of the third and six attributes (age and no. of children) will be done by the proposed interchanging: the first transaction interchanged by the last one, the second transaction by the before last one, and so on until the middle of both (if the no. of transactions were even then there is no problem there are two middle and will be interchanged, but if the no. was odd the middle will be still in its place).
4. The protection of the fifth attribute (marriage state) will be done by the proposed concatenation: if the state was marriage will take class I, if was divisor will take class II, if was worth will take class III, if was not marriage will take class IIII.

Mining results:

From applying the association rule data mining algorithm on the protected bank database, see fig. (3), will obtain the following rules:

If (E=I) and (F= less than 2) and (D= Ofecr4VLSvMOm1zVpR)
then (G= more than uv31WsfzM7U) and (I=more than FfNU7VGnk8E=)

So the above association rule (extracted knowledge) has a privacy because it is not understood by the group of mining and other users.

Analyzing results:

Now the database administrators will analyze the extracted knowledge by decryption the encrypted fields and return the values modified to its original values. The previous knowledge will mean the following:

If (the person marraige) and (have less than 2 child) and (live in Baghdad/rusafa)
then (will has more than 67$ as income) and (the income will be more than 4555$)

2.2. The Distributing Algorithm.

The proposed research suggest distributing the data as horizontal data distribution and vertical data distribution. Horizontal distribution refers to those cases where different database records reside in different places, see fig. (5), while vertical data distribution, refers to the cases where all the values for different attributes reside in different places, see fig. (6).

Horizontally:

In a horizontally distributed database, the transactions are distributed among n sites. The global support count of an itemset is the sum of all the local support counts. An itemset X is globally supported if the global support count of X is bigger than the given proposed support s% of the total transaction database size. A k-itemset is called a globally large k-itemset if it is globally supported.

Mining results:

From applying the association rule data mining algorithm on the horizontal distributed web log in/out database, see figure (5 a and b), will obtain the following:

From site (1) the miner group knows only the frequent itemsets of site 1, the same in site (2) so the privacy of the universal database is protected. In figure (5 a and b) the frequent itemsets are extracted by the administrators as in the previous section and the following association rule is an example:

If (A=122.22.3.18) and (B= 80) and (E=tcp) and (G=2:30)
Then (C=33.56.233.77)

Vertically:

Mining private association rules from vertically partitioned data, where the items are distributed and each itemset is split between sites, that done by finding the support count of an itemset. If the support count of such an itemset can be securely computed, then can check if the support is greater than the threshold, and decide whether the itemset is frequent. The key element for computing the support count of an itemset is to compute the scalar product of the vectors representing the sub-itemsets in the parties. Thus, if the scalar product can be securely computed, the support count can also be computed. The algorithm that computes the scalar product, as an algebraic solution that hides true values by placing them in equations masked with random values, is described in. The security of the scalar product protocol is based on the inability of either side to solve k equations in more than k unknowns. Some of the unknowns are randomly chosen, and can safely be assumed as private.

Mining results:

From applying the association rule data mining algorithm on the vertical distributed web log in/out database, see figure (6 a and b), will obtain the following:

From site (1) the miner group know only the frequent itemsets of site 1, the same in site (2) so the privacy of the universal database is protected. In figure (6 a and b) the frequent itemsets are extracted by the administrators as in the previous section and the following association rule is an example:

If (A=122.22.3.18) and (B= 80) and (E=tcp) and (G=2:30)
Then (C=33.56.233.77)

3- The Implementation:

To demonstrate the idea of this research in more clearly we would present the real implementation of that system as follow:

The first part of the implementation will concentrate on the privacy preserving by using modifications and encryption approaches which were explained in details in section (2.1), see figure (7).

In figure (7), see in the upper part of it the bank database with original data, down of it there is a command button called (modification and encryption), when it pressed then the program will display the bank database with modified and encrypted data. Then the encrypted secure database will be sent to the mining group. After arrival of the secure database to the miners then the miners will mine it and record and display the resulted encrypted association rules, see figure (8).

After recording (saving in a file) the encrypted association rules, then will be returned to the bank database administrators. The administrators will decrypt the association rules and begin to analyze them to extract the novel knowledge which support the performance of the applications related to their huge databases, see figure (9).

The second part of the implementation will concentrate on the privacy preserving by using distributing (fragmentation) approach which was explained in details in section (3.2), see figure (10).

In figure (10), display the original web log database, then the first command which is called (fragmentation) will display small interface to give the ability for administrators to select one of the two types of the distributing (horizontal and vertical), see figure (11).

Then after selecting for example the vertical fragmentation the program will split the original database into two databases since the no. of attribute is few. After sending these two parts to two group miners will get from them two files. Each file consists of the frequent itemsets of its part. The original database administrators will implement an algorithm to combine the two files of the frequent itemsets and extract the final frequent itemsets of the original complete database. Finally applying association rule algorithm on the final frequent itemsets to extract the novel knowledge, also see figure (10).

4. Conclusions.

In context with the results of the present study it can be concluded that:

- The work presented here, indicates the importance of interest of researchers in the area of securing sensitive data and knowledge from malicious users.
- The results of DM such as association rules, classes, clusters, …etc, will not be readily available for all team of work if take in our care the importance of support the DM algorithm with data protection for keeping the privacy of sensitive data.
- The proposed modification technique and using an encryption algorithm for protecting the data represent good solution to provide the privacy of data mining, since the modification and encryption are the main techniques in securing the sensitive of the data in all the fields of application.
- Selecting the twofish encryption algorithm keep the time consuming, since deal with static environment not dynamic such the internet so the symmetric encryption represent the optimized solution. Especially twofish because it still hardly to be broken in a little time.

TID	Local IP (A)	Local port (B)	Remote IP (C)	Remote port (D)	State (E)	Type (F)	Time stamp (G)
1	122.22.3.18	139	33.56.233.77	80	Listen	Tcp	2:30
2	122.22.3.18	139	44.56.78.22	50	Listen	Udp	5:50
m.	202.22.11.45	80	66.23.200.11	60	Listen	Udp	1:30
.							
.							
n.	122.22.3.18	90	22.22.3.198	77	Listen	TCP	4:23

Figure 1 the bank database.

ID	Identifier	Name	age	address	Marriage state	No.of child	income	Credit info.	Account
1	500	Suzan	30	Baghdad /rusafa/505/62/15	Marriage	3	1000	23456676	4555
2 . .	501 \	Ahmad	35	Baghdad /kargh/401/22/12	Devisor	6	67	23456677	54432987

Figure 2 the database for the web login/out file.

TI D	A	B	C	D	E	F	G	H	I
1	473368	WAG aA6T gTlk=	30	Ofecr4 VLSvM Om1zV pR	I	3	6zly qmw c8k Y=	EBOc x76Kc WJVn OBkj	FfNU 7VGn k8E=
2 . . .	373244	OMT G/WP r0qw=	35	Ofecr4 VLSvM YNfqO okZ	II	6	uv31 Wsf zM7 U=	WPqe 56mj NTV VnOB	8f9pS HWX WC9 VnOB kjC

Figure (3) the bank database after protect all its data, keeping the privacy of the person information.

Figure (4): the implementation of twofish encryption algorithm.

TID	(A)	(B)	(C)	(D)	(E)	(F)	(G)
1	122.22.3.18	139	33.56.233.77	80	Listen	Tcp	2:30
2	122.22.3.18	139	44.56.78.22	50	Listen	Udp	5:50
.							
.							

Figure (5): a: horizontal distribution site (1)

TID	(A)	(B)	(C)	(D)	(E)	(F)	(G)
m.	202.22.11.45	80	66.23.200.11	60	Listen	Udp	1:30
.							
.							
n	122.22.3.18	90	22.22.3.198	77	Listen	TCP	4:23
.							

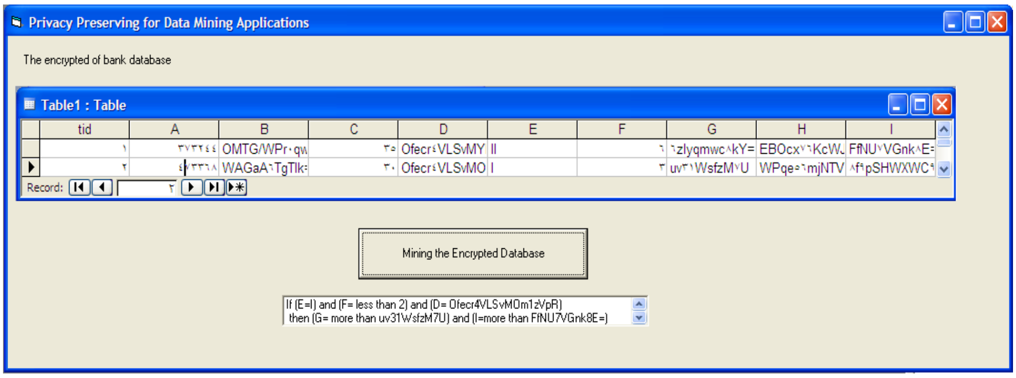
Figure (5): b: horizontal distribution site (2)

TID	(A)	(B)	(C)
1	122.22.3.18	139	33.56.233.77
2	122.22.3.18	139	44.56.78.22
m.	202.22.11.45	80	66.23.200.11
.			
.			
n.	122.22.3.18	90	22.22.3.198

Figure (6): a: vertical distribution site (1)

TID	(D)	(E)	(F)	(G)
1	80	Listen	Tcp	2:30
2	50	Listen	Udp	5:50
m.	60	Listen	Udp	1:30
.				
.				
n.	77	Listen	TCP	4:23

Figure (6): b: vertical distribution site (2)



Privacy Preserving for Data Mining Applications

The encrypted of bank database

Table1 : Table

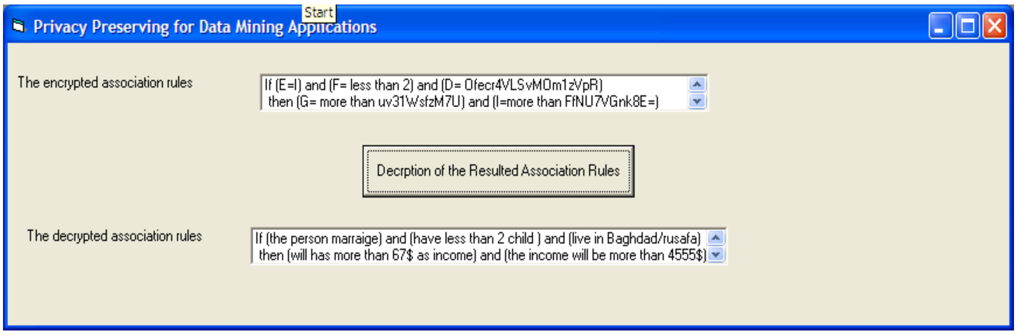
tid	A	B	C	D	E	F	G	H	I
١	٣٧٣٢٤٤	OMTG/WPr٠qw	٣٥	Ofecr٤VLSvMY	II	٦	١zlyqmwc٨kY=	EBOcx٧١KcWJ	FfNU٧VGnk٨E=
٢	٤٧٣٣٦٨	WAGaA١TgTlk=	٣٠	Ofecr٤VLSvMO	I	٣	uv٣١WsfzM٧U	WPqe=١mjNTV	٨f٩pSHWXWC٩

Record: ٢

Mining the Encrypted Database

If (E=I) and (F= less than 2) and (D= Ofecr4VLSvMOm1zVpR)
then (G= more than uv31WsfzM7U) and (I=more than FfNU7VGnk8E=)

Figure (7): The implementation of mining the secure database by association rule algorithm.



Privacy Preserving for Data Mining Applications

Start

The encrypted association rules

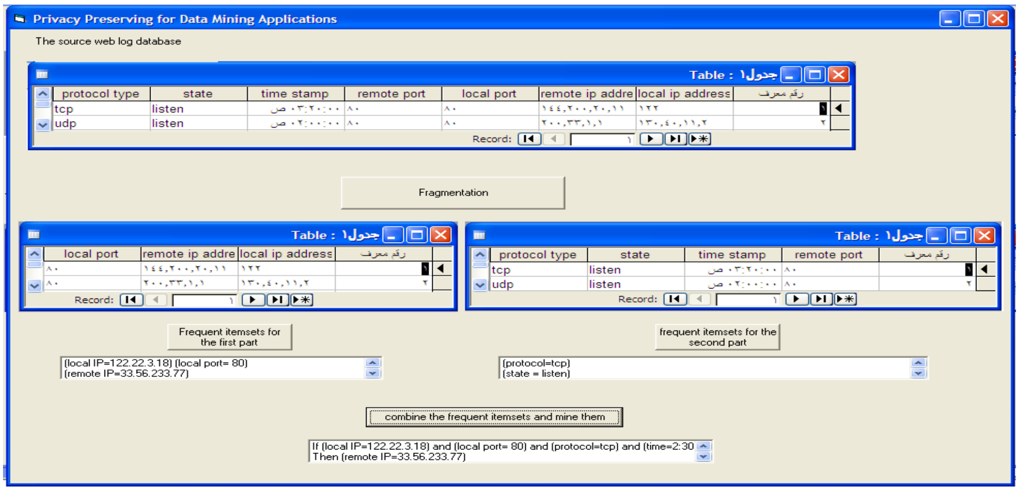
If (E=I) and (F= less than 2) and (D= Ofecr4VLSvMOm1zVpR)
then (G= more than uv31WsfzM7U) and (I=more than FfNU7VGnk8E=)

Decrption of the Resulted Association Rules

The decrypted association rules

If (the person marraige) and (have less than 2 child) and (live in Baghdad/rusafa)
then (will has more than 67$ as income) and (the income will be more than 4555$)

Figure (8): The decryption of mined encrypted association rules.



Privacy Preserving for Data Mining Applications

The source web log database

Table : جدول١

protocol type	state	time stamp	remote port	local port	remote ip addre	local ip address	رقم معرف
tcp	listen	٠٣:٢٠:٠٠ ص	٨٠	٨٠	١٤٤,٢٠٠,٢٠,١١	١٢٢	١
udp	listen	٠٢:٠٠:٠٠ ص	٨٠	٨٠	٢٠٠,٣٣,١,١	١٣٠,٤٠,١١,٢	٢

Record: ١

Fragmentation

Table : جدول١

local port	remote ip addre	local ip address	رقم معرف
٨٠	١٤٤,٢٠٠,٢٠,١١	١٢٢	١
٨٠	٢٠٠,٣٣,١,١	١٣٠,٤٠,١١,٢	٢

Record: ١

Table : جدول١

protocol type	state	time stamp	remote port	رقم معرف
tcp	listen	٠٣:٢٠:٠٠ ص	٨٠	١
udp	listen	٠٢:٠٠:٠٠ ص	٨٠	٢

Record: ١

Frequent itemsets for the first part

(local IP=122.22.3.18) (local port= 80)
(remote IP=33.56.233.77)

frequent itemsets for the second part

(protocol=tcp)
(state = listen)

combine the frequent itemsets and mine them

If (local IP=122.22.3.18) and (local port= 80) and (protocol=tcp) and (time=2:30
Then (remote IP=33.56.233.77)

Figure (9): privacy preserving by distributing (fragmentation) approach.

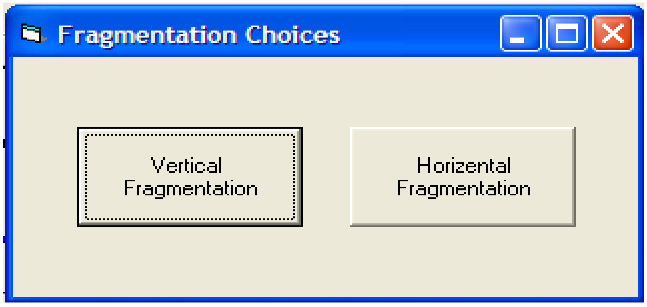


Figure (10): Interface which will make the administrators able to select one of the two types of the distributing**.**

References:

1- . Chen M. S, Han J., and Yu. P. S. **“Data mining: An overview from a database perspective”**. *IEEE Trans. Knowledge and Data Engineering*, 8:866-883, 1996.

2- Dunham Margraet H., ***"Data mining: Introductionary and Advanced Topics"***, *Southern Methodist University*, 2003.

3- Fayyad U. M. , G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. **“Advances in Knowledge Discovery and Data Mining”**. *AAAI/MIT Press*, 1996.

4- Han J. and M. Kamber. **“Data Mining: Concepts and Techniques”**. *Morgan Kaufmann, 2000.*

5- Kantardzic M., ***“DM Concepts, Models, Methods and Algorithms”,*** *John Wiley &Sons, 2003.*

6- www.twocrows.com, *technology report*, ***"Introduction to Data Mining and Knowledge Discovery"***, Third Edition, Two Crows Corporation, 2000.