International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 2 Issue: 5 1332 – 1337

Collusion Avoidance in Fingerprinting Outsourced Relational Databases with Knowledge Preservation

Ms. Varsha Waghmode
Computer Engineering Department
DYPSOET, Lohegaon
Pune, India
waghmodevarsha@gmail.com

Ms. A.A. Mohanpurkar
Computer Engineering Department
DYPSOET, Lohegaon
Pune, India
yasharti@gmail.com

Abstract- Large databases are mined to gain knowledge from them. Data is also mined to help decision makers to make efficient decisions. But not every organization which is collecting data can mine the data e.g. Hospitals. So there arises need to handover this data to other mining expert organizations that can discover the knowledge within it. But in today's world one has to be careful about protection of data in terms of ownership theft, tampering and various other kinds of attacks. Fingerprinting for relational databases has emerged as a compound solution which provides such protection to relational data. Although fingerprinting provides security, the challenge is that after inserting fingerprint, the data must remain useful for the intended purpose i.e. insertion of marks should not reduce the usability of the original data. We have proposed system which addresses this issue by providing automated usability constraint model [1]. Usability constraint model provides the distortion band for each feature within which the values can change [1]. Also proposed fingerprinting technique is used to provide right protection on database as well as it helps to indentify traitor (if any). The proposed system has new insertion and detection algorithm which is based on hashing technique. This insertion algorithm avoids collusion as well as it reduces the fingerprint insertion and detection time to a large extent.

Keywords-Collusion Avoidance, Fingerprinting, Knowledge preserving, Ownership protection, Relational database, Usability constraint

*****

I. Introduction

The databases need to be mined to extract the knowledge from them and the same purpose leads owners (of database) to outsource them to third party knowledge miners such as data experts, researchers etc. Thus relational databases need to be shared between researchers and owners which bring forth different issues of ownership protection and illegal redistribution etc. Watermarking is well known technique to impose and to prove ownership on relational data [1][2][7][9][10] while fingerprinting is another well-known technique used to identify the recipient to whom the data has been provided(Traitor identification) [3][4][5][6][12].

In case of numeric relational databases inserting fingerprint induces some modifications in data which may cause loss of knowledge in them or may yield wrong results. Consider the case of medical database. The health maintenance organization use KDD techniques and historical data of patients to determine which of its enrollees may be at risk for certain diseases [14]. The changes (due to fingerprint insertion) in data may cause loss of knowledge and thus may result into misdiagnosis [6] [2]. Such misdiagnosis puts enrollee's life at risk or may lead to increased cost of health care. Thus fingerprinting should not modify data to an extent where it may result in loss of knowledge [2][13].

M. Kamran and Muddassar Farooq [1] proposed a usability constraint model to find allowable alteration in each feature and also a watermark insertion and detection algorithm. Watermarking technique can prove only ownership. Our system extends the M. Kamran's work by fingerprinting databases. Fingerprints can not only prove the ownership protection but it can also be used for traitor identification. Our proposed system reduces the time complexity of insertion and detection over the M. Kamran's system by large extent and still robustness is maintained and traitor identification is possible.

In the fingerprinting era the important problem is to avoid collusions. A lot research has been done for collusion secure fingerprints [3] [15]. Also Li [5] has proposed a fingerprint insertion using pseudorandom sequence generator which uses primary key of database. We have proposed an insertion algorithm based on hashing technique which inserts the fingerprint for each buyer in such way that it leads to collusion avoidance. The insertion technique is primary key independent and detection algorithm efficiently identifies the traitor. The proposed system is robust against tuple deletion, tuple insertion attacks.

Here knowledge preservation is achieved using Kamran's automated usability constraint model and then fingerprinting of database. As a result, the classification accuracy of the dataset remains unaltered. In addition to this, the inserted

fingerprint remains imperceptible and robust against any type of sophisticated attacks that can be launched on the dataset.

II.MOTIVATION

Fingerprinting relational database is a challenge as in case of numeric features it significantly changes data. There are many databases like medical records, weather databases where numeric attributes play important role. The proposed system aims to provide security by fingerprinting numeric attributes and preserve the knowledge by applying Usability constraint model. Consideration of Local and global constraints in the model helps to optimize the amount of acceptable alteration of numeric attribute while fingerprinting. Fingerprinting technique helps to trace and identify the traitor (if any).

III. RELATED WORK

M. Kamran and Muddassar Farooq [1] proposed formal usability constraint model. This model is applied as an input to watermarking algorithm in automated manner. This model is the one that facilitates a data owner to define usability constraints to preserve the knowledge contained in the dataset. Ownership protection is achieved without loss of knowledge but the time complexity of insertion and detection algorithm is found to be high.

M. Kamran and Muddassar Farooq [2] presented information preserving watermarking technique based on classification potential of the feature. This system mainly aims at preserving knowledge in the EMR (Electronic Medical record) as change in information of EMR might not only result in a life threatening scenario but also might lead to significant costs for treatment to the patients. The proposed system works not only for EMR but for any numeric relational database.

Yingjiu Li [5] presented a marking scheme that permits an arbitrary mark bit-string to be embedded in a relation using a single secret key. The mark bit-string can be used to represent different buyers who purchase the database. The detection algorithm tests whether a key was used to mark a relation and, if so, it returns the actual mark bit-string that was embedded. The marking scheme can be used for both watermarking and fingerprinting. The only difference is that in watermarking same bit-string is embedded and detected but in fingerprinting different bit-strings are embedded as well as detected. This marking scheme depends on primary key. The proposed fingerprint insertion technique is independent of primary key.

Julien Lafaye [6], proposed an optimized fingerprinting system for databases under constraints. It features built in usability constraints definition language. The user has to explicitly specify acceptable change in each feature of every database. But in proposed system the automated usability constraint model does it implicitly.

Ersin Uzun and Bryan Stephenson [9], proposes security of relational databases using fingerprinting and watermarking for business outsourcing. In business process outsourcing the change in numerical values is not acceptable, such change may disrupt business process. The proposed system is interested in preserving the knowledge instead of data. In proposed system changes in data to some acceptable extent is allowed.

Dan Boneh and James Shaw [3] discuss the method for assigning code words for the purpose of fingerprinting digital data. This method is efficient in terms of number of users and coalition size. The proposed system uses randomness property to construct the fingerprint code.

Hans George Schaathun [4] improved Boneh-Shaw code. Here it is said that either inner code or outer code may be replaced to reduced code length. The proposed algorithm help to avoid collusion by inserting mark at random places in database. The sequence of marks inserted for single table is different for each buyer. So this technique helps to avoid collusion

IV. PROPOSED SYSTEM

The proposed system is about protecting the numeric relational databases against ownership attack and illegal redistribution. Fingerprinting is a technique used to achieve the same. But the main concern of the system to preserve knowledge within database as inserting a fingerprint mark may tend to make changes in data values, which may result in loss of knowledge. Whenever data is mined after applying fingerprint it should return same knowledge as it was there before fingerprinting. As well as the applied fingerprint should be able to detect the traitor efficiently and should resist collusions. So the focus of proposed system is to preserve the knowledge as well as apply a collusion resistance fingerprinting and trace traitor (if any). The proposed system achieves this goal in three steps:

1. Usability constraint model: This is automated model for defining usability constraints on database. This model is propose by M.Kamran [1]

 ISSN: 2321-8169
1332 – 1337

2. Fingerprinting: Fingerprinting not only provides ownership protection but also helps to identify the traitor. Proposed system uses Boneh-Shaw's collusion secure code with some variations.

3. Verification system: This part of system is used to check whether knowledge is preserved or not. This system uses different types of classification algorithm like NaiveBayes, SMO, IBK, Bagging, JRip, and J48 to test the result. The results are shown using classification statistics.

V. KNOWLEDGE PRESERVING FINGERPRINTING SCHEME

The Proposed system is divided into three parts. First it develops usability constraint model which is given as input to fingerprint insertion algorithm [1].

The fingerprint part contains fingerprint encoding, decoding and algorithm for tracing the traitor. At the end proposed system verifies the main results to check the knowledge preservation.

Fig. 1 shows Architecture of proposed system.

A. Usability Constraint model

The knowledge is preserved in database if it meets following three constraints [1].

1. The class label of every tuple should remain same before and after fingerprinting. If S_O is class label of tuple before Fingerprinting and S_{WF} is class label after fingerprint then,

$$S_O = S_{WF} \quad (1)$$

2. Classification potential of each feature should remain same before and after fingerprint. Let CPT_O ,CPT_{WF} classification potential of original database and fingerprinted databases respectively then,

$$CPT_O = CPT_{WF} \quad (2)$$

3. The Distribution of data for every feature should remain same before and after fingerprint. Let H_O ,H_{WF} is distribution of data before and after fingerprint respectively for feature x then,

$$H_O = H_{WF} \quad (3)$$

To meet these requirements Kamran [1] has suggested the constraint model which defines two types of constraints Local usability constraint and global usability constraint.

i. Local Usability Constraint

These constraints are defined by calculating mutual information of feature. Local constraint is a tuple of mutual information for each feature in group g_i,

$$L_{CONSTRAINT} = I(F)$$

Local constraints are applied to every group. Local constraints help to optimize the acceptable alteration in every feature. They also help to meet constraint 2 and 3 in knowledge preserving constraint model

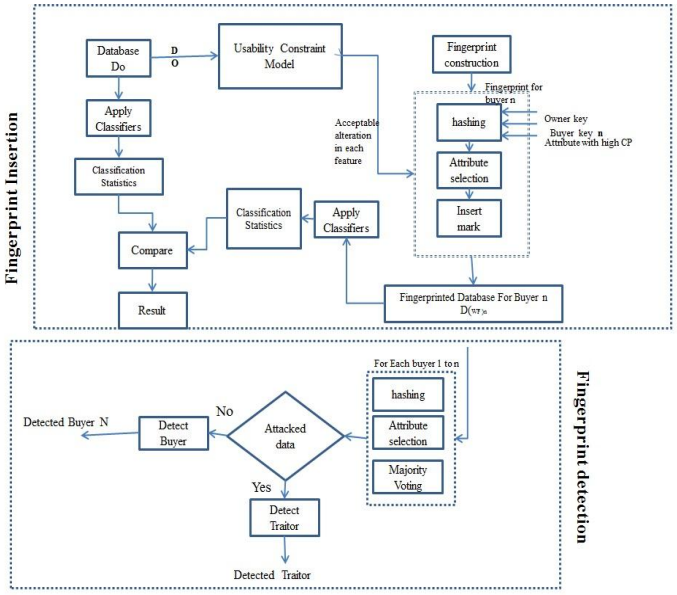

Fig.1 Architecture of proposed System

ii. Global Usability Constraint

Global usability constraint is defined using five well known feature selection scheme (Information Gain, Information Gain Ration, CFS, CBF, and PC). Global constraints are applied at group level as well as global database level.

$$G_{CONSTRAINT} = (IG, IG_r, CFS, CBF, PCA)$$

Global constraints help to meet constraint 1 and 2 in knowledge preserving constraint model.

B.Fingerprinting encoding, decoding and traitor identification

Fingerprinting is used to mark the database with unique buyer specific identification marks. Different Buyers are marked with different identification marks. These marks help to identify owner and buyer of database and in case of illegal redistribution such marks help to identify traitor

Algorithm1. Insertion of fingerprint in database for buyer n

Input: Original dataset Do Owner's secret key K_W, Buyer's ID K_B and Acceptable alteration in each feature Δ, Fingerprint code F
Output: Fingerprinted Database D_{WF} for buyer n, Alteration table

```
Temp==Do
FOR each row r
   Attribute =HASH (Kw, KB, Attribute with high CP)
   IF (Attribute not equal to Attribute with high CP)
          IF F (bit) ==1
          Attribute (value) = Attribute (value) + Attribute (Δ)
          Alteration table= Attribute (Δ)
          ELSE
          Attribute (value) = Attribute (value) + Attribute (Δ)
          Alteration table= Attribute (Δ)
          END IF
  END IF
END FOR
return DWF, Alteration table
```

i. Fingerprint construction:

In our proposed scheme the fingerprint code can be constructed using any technique. The proposed scheme uses Tardo's code. But the code can be constructed using any method. The proposed technique avoids collusion using a typical insertion scheme, hence the fingerprinting code need not to be collusion secure i.e. can be any bit stream.

ii. Fingerprint insertion:

Algorithm 1 shows fingerprint insertion. It uses hashing technique. The hash value H(row) calculated for each row using owner's secret key, Buyers secret key and value of attribute with high classification potential. For each buyer different hash value sequences are generated. This identifies the attribute within which the mark will be inserted. If the fingerprint bit is 0 the tolerable alteration is subtracted from attribute value and if bit is 1 tolerable alteration is added to attribute value. The insertion algorithm is shown in figure. The proposed insertion algorithm reduces insertion complexity to large extent.

iii. Fingerprint detection:

Algorithm 2 shows fingerprint detection.

The acceptable alteration Val in each feature is calculated using usability constraint model. This Val is compared with alteration table value.

Algorithm2: Fingerprint Detection

Input: Fingerprinted Database DWF ', Alteration table
Output: Buyer F(n)

```
Matching threshold=70%
  Checking threshold=50%
One=0;
Zero=0;
Row=row/2
For each buyer 1 to n
   For each Row
          Attribute =HASH (Kw, KB, Attribute with high CP)
          Val=Attribute (Δ)
          If alteration>Val
                Then F'(bit)==1
                One++
          End if
          If alteration<Val
                Then F'(bit)==1
                Zero++
          End if
   End for
Apply majority voting to get Fingerprint F'
Match: F'==F(Buyer)
   If Match> = Checking threshold
          Continue to detect F'
          Row=total no of rows
   Else if Match>=Matching threshold
          Detected a buyer i          iє1 to n
          stop
   Else
          Buyer++
End for
Return Buyer F(n)
```

Checking threshold: The system will continue to detect F for particular buyer till this threshold. If it is false then we move to check for next buyer

Acceptance threshold: For any buyer if checking threshold is true then the system will continue to detect F till acceptance threshold.

Once for any buyer we reach at Acceptance threshold we have detected the buyer ID.

Majority voting is applied to confirm the fingerprint code.

 ISSN: 2321-8169
1332 – 1337

iv. Traitor Tracing.

To trace the traitor same algorithm as detection is used on the attacked database. The detected fingerprint is checked with each buyer to get the traitor.

C. Verification system. (Analyzing the results)

This part of proposed system verifies whether the applied fingerprinting is lossless.

Let CST_O and CST_{WF} are classification statistics on database before and after fingerprinting.

If $CST_O = CST_{WF}$ Then $S_O = S_{WF}$, $CPT_O = CPT_{WF}$, $H_O = H_{WF}$ which in turn means if classification statistics on database before and after fingerprinting remain same then knowledge is preserved[1].

The information loss is defined as

$$CST_{Loss} = \frac{|CST_O - CST_{WF}|}{CST_O} * 100$$

The knowledge is preserved [1] if

$$CST_{Loss} = 0$$

So this verification system checks whether the knowledge is preserved or not. We can also calculate percentage of loss in knowledge (if any) using above formula.

VI. IMPLEMENTATION

All Experimentation is performed using Pentium processor and 2 GB RAM. The operating system is windows 7(32 bit) with JDK1.5 and Net beans IDE7.1.0 The experimentation is performed on Breast tissue database obtained from UCI repositories.

VII.RESULTS

The proposed system should preserve the knowledge in database. Fingerprint insertion changes the data but information loss is zero. To prove this we will calculate classification statistics on database before and after fingerprint. Let CSTO be the statistics before fingerprint and CSTWF be the statistics after fingerprint,

$$CST_O = \{TP_{rate}, Fprate, Rb\}$$

$$CST_{WF} = \{(TP_{rate})_{WF}, (FP_{rate})_{WF}, Rb_{WF}\}$$

Then the expected result is,

$$CST_O = CST_{WF} \quad (4)$$

Change in statistics before and after fingerprint	Do TPrate	DWF TPrate	Δ TPrate	Do FPrate	DWF FPrate	Δ FPrate
NaiveBayes	0.708	0.698	0.01	0.054	0.056	0.02
SMO	0.604	0.604	0	0.082	0.082	0
IBK	0.717	0.717	0	0.056	0.056	0
Bagging	0.726	0.726	0	0.055	0.055	0
JRip	0.632	0.651	0.019	0.077	0.077	0
J48	0.66	0.66	0	0.068	0.068	0

Table 1

Table I shows Effect of fingerprint insertion after applying usability constraint on TPrate and FPrate. Different Learning algorithm is used to classify the original dataset and fingerprinted dataset. The table shows difference between TPrate and FPrate.

M.Kamran has inserted watermark into dataset but our system has extended his work to insert fingerprints. Watermark can only be used for ownership protection but fingerprinting is used not only for ownership protection but also to trace traitor. As well as insertion complexity of Kamran technique is too high.

Let, N=No of Rows=107, M=No of columns=9,
Length_Fingerprint=1000, noof buyers=5

By Kamran's Method

Complexity=N * M* Length_Fingerprint* noof buyers

4815000 =107*9*1000*5

The Proposed System

Complexity=N * N* Length_Fingerprint* noof buyers

535000 =107*1000*5

Thus it reduces the complexity to larger extent

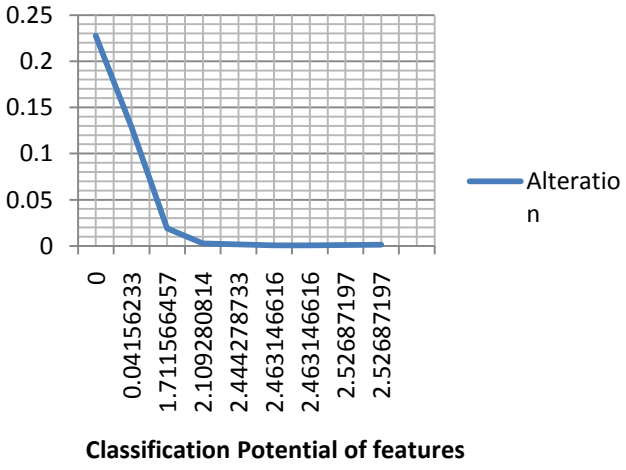

Fig 2: Graph of Classification potential Vs acceptable alteration

Fig 2 shows a graph of classification potential verses acceptable alteration. Feature with high Classification potential shows less amount of acceptable alteration (near to zero) where as features with low classification potential shows more amount of alteration. In graph horizontal axis is classification Potential of features and vertical axis shows amount of acceptable alteration

A. Collusion Avoidance

The proposed system uses the hash function for insertion. The hash function is used to identify the attribute for each row where it inserts a fingerprint mark. Using this technique we generate different sequence of attribute to be marked for each buyer. As a result the pattern of insertion is randomized for each buyer. Hence the fingerprint copies for different buyers of same databases are so different that they cannot collude to find the places of insertion of fingerprint marks.

B. Robustness against different attacks:

Tuple deletion attack: If more than 50% tuples are deleted by an attacker still we can detect the fingerprint. We can detect the fingerprint from any row of table.

Tuple Addition attack: Adding of some tuples to the dataset does not degrade the detection of the bits; still the system is able to detect the fingerprint efficiently.

Majority voting is confirms the detection of the correct fingerprint bits.

VIII. CONCLUSION

Fingerprinting provides security against the ownership theft as well as it also helps trace the traitor if any unauthorized copy is found. In case of relational databases, insertion of fingerprint bit can change numeric data to some extent. This change in numeric data may lead to loss in knowledge from database. M.Kamran's [1] usability constraint model is used to find acceptable alteration. We have extended M.Kamran's [1] work by inserting a fingerprint on database while achieving a knowledge preservation. We have achieved collusion avoidance by using hashing technique. Also we have reduced the complexity of insertion by 70% over M.Kamran's insertion technique. The proposed insertion algorithm is independent primary key and it can efficiently trace the traitor.

REFERENCES

[1] M. Kamran and Muddassar Farooq:"A Formal Usability Constraints Model for Watermarking of Outsourced Datasets", IEEE transactions on information forensics and security, Vol . 8, no. 6, June 2013

[2] M. Kamran and Muddassar Farooq, "An Information-Preserving Watermarking Scheme for Right Protection of EMR Systems", IEEE Transactions on Knowledge and Data Engineering, Vol. 24, No. 11, November 2012.

[3] Dan Boneh and James Shaw," Collusion Secure Fingerprinting For Digital data", IEEE Transaction on Information Theory, Vol. 44,No. 5,September 1998.

[4] Hans George Schaathun,"The Boneh-Shaw Fingerprinting is Better Than We Thought", IEEE Transaction on Information Forensics and Security Vol 1. No 2.June 2006.

[5] Yingjiu Li, "Fingerprinting Relational Databases: Schemes and Specialties", IEEE Transactions On Dependable And Secure Computing, Vol. 2, NO. 1, January-March 2005.

[6] Julien Lafaye, David Gross-Amblard, Camelia Constantin, and Guerrouani "Watermill: An Optimized Fingerprinting System for Databases under Constraints" IEEE Transactions on Knowledge and Data Engineering, vol.20, No.4, APRIL 2008

[7] Raju Halder, Shantanu Pal, Agostino Cortesi, "Watermarking Techniques for Relational Databases: Survey, Classification and Comparison", Journal of Universal Computer Science, Vol. 16, no. 21 (2010).

[8] Sunita Beniwal, Jitender Arora "Classification and Feature Selection Techniques in Data Mining" International Journal of Engineering Research & Technology (IJERT), Vol. 1 Issue 6 August – 2012, ISSN: 2278-0181

[9] Ersin Uzun and Bryan Stephenson, "Security of Relational Databases in Business Outsourcing" , HP Laboratories, HPL-2008-168

[10] Radu Sion , Mikhail Atallah , Sunil Prabhakar "Rights Protection for Relational Data" IEEE Transactions on Knowledge and Data Engineering, Vol. 16, No. 06, June 2004

[11] Rakesh Agrawal, Peter J Haas, Jerry Kiernan, "Watermarking relational data: framework, algorithm and analysis", The VLDB Journal(2003)/ Digital object identifier(DOI) 10.1007/s00778-003-0097-x.

[12] Alexander Barg, G. R. Blakley, and Grigory A. Kabatiansky, "Digital Fingerprinting Codes: Problem Statements, Constructions, Identification of Traitors", IEEE Transaction on Information Theory vol.49, No. 4, April 2003.

[13] Ling Qiu, Yingjiu Li, Xinatao Wu,"Protecting business intelligence and customer privacy while outsourcing data mining tasks" Knowledge Information System(2008)17:99-12,DOI 10.1007/s10115-007-0113-3

[14] Kristin B. DeGruy, MSHS,"Healthcare Applications of Knowledge Discovery in Databases".

[15] Tardos, G., "Optimal probabilistic fingerprint codes," J. ACM 55(2), 1–24 (2008).