International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

Enhanced Robust and Reversible Watermarking For Supervised Learning Data

Madhuri V. Gaikwad[1], Prof. R. A. Kudale[2]

Dept. of Computer Engineering SKNCOE, Savitribai Phule Pune University, Pune, India

Assistant Professor, Dept. of Computer, Engineering, SKNCOE, Savitribai Phule, Pune University, Pune, India

Abstract: ***Large database have large data structure and have their large relational data for sharing for particular target or authenticate target but there is some other parties attack on data easily and use that data illegally. Attacker can gain ownership on that sharing data. While original data get modified and quality of data also degraded so this original data not useful for any extraction information system, it gives wrong data or reduced data. For that we used a system Reversible Watermarking which protects data from Attack of middle parties while sharing data and also preserve ownership of the data .quality of the data also preserve avoid the data tampering. Feature selection in RRW uses all combinations of features to calculate importance of the features. In in supervised learning, feature's importance depends on co-relation between Feature and class variable, there is no need to consider all combination in such case. Also RRW does not support non-numeric data. We introduced technique which works on nominal data and uses less features for calculation which enhance the speed and accuracy and performance of RRW.***

Keywords: Reversible watermarking; genetic algorithm; data recovery; data quality.

1. Introduction

Most of the important information systems are based on Relation database and these information systems are used to gain knowledge hidden inside this large data. Many times this relational database is used by many parties to extract information collaboratively. Sharing such important relational databases makes becomes easy target for attacker which uses this data illegally. This attacker may claim ownership on the shared data therefore data ownership preservation technique is needed.

Watermarking methods are used in the literature for ownership preservation and preserve data tampering. This watermarking techniques are applied widely on all types of the data i.e. images, audio, video, databases. Basic problem of watermarking is when watermark i.e. ownership information is embedded into the original data, original data gets modified. Data modification due to watermarking may cause quality degradation and this data may not be useful for information extraction or data mining operation. Utility of the data is reduced means data gives wrong information extraction result.

Reversible watermarking is the solution for above problem which maintains the data quality by data recovery option with ownership preservation. Fingerprinting, data hashing, serial codes are some other techniques used for ownership protection.

Reversible watermarking objective are:
1) Data ownership protection
2) Reduce data quality degradation due to watermark embedding

These objectives are achieved by data recovery option in which original data can be recovered by removing the watermark information from watermarked data. Many times this reversible watermarked data suffers from distinct data tampering attack.
1) Data insertion attack
2) Data deletion attack
3) Data modification attack.

In paper [1], RRW is reversible watermarking scheme is applied which gives solution for all above requirement
1) Data ownership protection
2) Data quality is maintained
3) Effective against malicious attacks like data insertion attack, data deletion attack, data Modification attack.

RRW used the most irrelevant feature for embedding watermark information. There is scope to apply this scheme to preserve data ownership of dataset which are used for supervised learning. In such data sets features importance depends on the relevance between the feature and the class feature. In RRW relevance of the feature is checked with all other features in the database but when we apply RRW to supervised learning database there is no need to calculate the relationship with all other features. RRW is not applicable to non-numeric features. We introduced a technique which does not require use of all the features instead we pruned the number of features and increase the speed and accuracy of the process. Our method also works on non-numeric data.

The subsequent sections of the paper are structured as follows: In Section 2, literature review is provided. In Section 3, system architecture is given. In Section4, mathematical model is given. In Section 5, experimental results are discussed. Finally, the paper is concluded in Section 6.

2. Literature Review

Paper [1] proposed irreversible watermarking specially tailored for relational databases. This paper addressed first

time there is need of watermarking for relational databases. Paper designed special watermarks which are suitable for relational database. Paper also addressed the possibility of attacks on watermarked data. Problem of this scheme is watermarking is irreversible.

Paper [2] provides the reversible feature to the watermarking i.e. original data recovery is possible from watermarked relational data. Watermark system exploit methods of arithmetic operations on numeric features and perform transformations. The watermark information is normally embedded in the LSB of features of relational databases to minimize distortions.

Technique proposed in this paper [3] minimizes the distortion in the data, increases watermarking capacity and reduces the false positive rate. This technique is used in recent watermarking algorithms. Problem of this technique is that robustness of the technique can be compromised on heavy attacks.

Paper [4] proposes one of the recent Prediction error expansion watermarking techniques (PEEW) which incorporates a as opposed to a difference operator to select candidate pixels or features for embedding of watermark information. The PEEW proposed technique by Farfoura and Horng is fragile against malicious attacks as the watermark information is embedded in the fractional part of numeric features only. In order to ensure integrity, detect malicious modification and protect ownership rights, paper [5] proposes a watermarking algorithm based on parameterized tuple partitioning and whitespaces, using a public watermark. The watermarking scheme is non-intrusive, resilient, blind, reversible and suitable for databases of any size with reasonable performance on embedding and extraction. Moreover, proposed method emphasize locatability of malicious modifications within the scope of predefined tuple sets, and support incremental watermarking to cope with the dynamic nature database systems are subject to.

Paper [6] proposes reversible watermarking technique which is robust against the malicious attacks like insertion attack, modification attack and deletion attack. RRW focuses on maintaining data quality so that data will be useful further for information extraction. Problem of this technique is it cannot be applied on non-numeric data.

3. System Architecture

3.1 Existing System

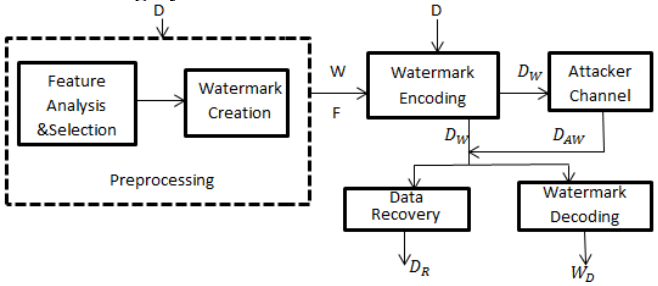

Figure 1: Architecture of Robust and Reversible Watermarking (RRW)

RRW works in following steps:

1) Preprocessing

To maintain the quality of the data RRW find the feature which is less important and embeds the watermarking data into that feature. In preprocessing step importance of the all features in the feature set of database is evaluated. Watermark should be in large size such that it preserves the data ownership easily and small enough that data quality is not degraded.

- **Feature analysis and selection:**

To find the importance of the features Mutual information (MI) is calculated. Mutual information of each feature with all other feature is calculated using following equation

$$MI(A, B) = \sum_a \sum_b P_{AB}(a, b) \log \frac{P_{AB}(a,b)}{P_{AB}(a) P_B(b)} \quad (1)$$

Where MI(A,B) measures degree of correlation of between A and B feature by marginal probability distribution and joint probability distribution.

To calculate the overall MI of feature summation of MI of feature with all features is taken

$$MI(A) = \sum_{i=0}^{n} MI(A, i) \quad (2)$$

After calculation of MI values of all features, features with MI values less than threshold values are considered for embedding watermarks

- **Watermark creation**

Watermark creation is done by using Genetic algorithm population-based computational model, basically inspired from genetic evolution. Initial random population of binary strings called chromosomes is generated. Gene values of each chromosome represent l-bit watermark string. In the proposed scheme, the GA is populated with a constrained fitness function to acquire an optimal change in data that will ensure data quality while embedding the watermark. Watermark creation problem is considered as multi objective optimization problem and solved using genetic algorithms. Watermarks are created and each watermarks beta value is calculated using constrained fitness function which is the measure of extent of change by watermark in original data. Data quality is ensured by imposing the following usability constraints λ on original and watermarked data in equation 3

$$\text{Mean (Dw)} - \text{Mean (D)} = 0$$
$$\text{Variance (Dw)} \_ \text{Variance (D)} = 0$$
$$\text{MI (Dw)} - \text{M(D)} = 0 \quad \ldots\ldots\ldots(3)$$

Where Dw is watermarked data and D is original data. Output of this phase is optimal chromosomal String (Watermark string with length l) and beta value which represents tolerable amount of change to embed in the feature values.

2) Watermark encoding phase

From previous steps system gets watermark bit string and Beta value for values of each feature. Watermark encoding is straight forward step

Input: D, W, B
Process:
For each w bit in the watermark

```
    For each r tuple value of selected feature
      If bit value ==0
       Then then Dw = Dr – B
        Changes are added in Vector V
         End If
      If bit value = 1
       Then Dw = Dr +B
        Changes are added in Vector V
       End IF
    End For
End For
Return Dw, V
Output: Dw, V
```

3) Watermark Decoding Phase

In this step watermark information is extracted from watermarked data Dw. Decoding phase consists of mainly two steps for each feature in Dw watermark bits are detected staring from least significant bit to Most significant bit. This process is carried out by using change matrix Nr. In second step bits of watermark are decoded according to percentage change values of watermarked data. If percentage change is less than zero then is detected watermark 1 else it is 0.

4) Data recovery Phase

In this phase, original data is recovered from watermarked data. To recover the data watermarked data, detected watermarked and beta values computed in step second are use. If detected watermark is 1 for given r value in R tuple in step 3 then respective beta value is added into r value to recover the data. If detected watermark is 0 for given value in R tuple in step 3 then beta value is subtracted from r value to recover the data.

```
Input: Dw, dtW , b
Process:
For r = 1 to R do
  For b = L to 1 do
    If dtW(r,b) ==1
     Then Dr = Dwr + B
   End If
   If dtW(r,b) ==0
    Then Dr = Dwr -B
   End If
  End for
End for
Output:
D
```

- **Disadvantage of the existing systems**

1) In feature selection phase, it calculates co relation of feature with all other features present in the database to find the importance of the feature but when we consider database / dataset of supervised learning training data there is no need to consider all other features.
2) Existing system does not work on non-numeric data.

3.2 Proposed System

Proposed system is extension of the RRW existing system which is designed to remove afore mentioned disadvantages and to inherit all good features of it.

- **Working of the Proposed system:**

Proposed system architecture and working varies only in two steps preprocessing and data recovery phase

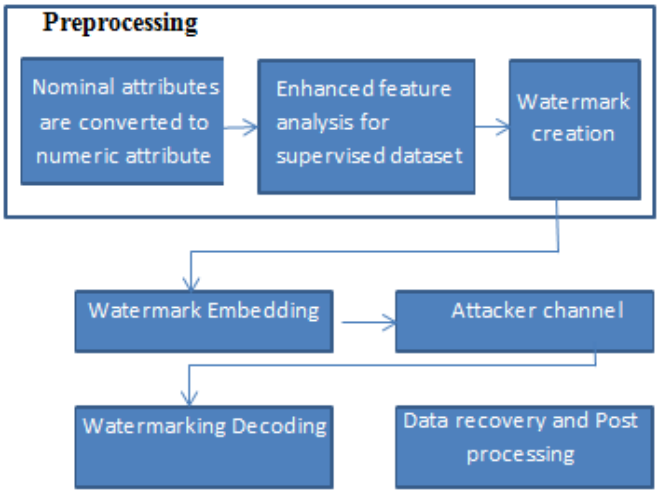

Figure 2: Architecture of Proposed system

1) Preprocessing Phase

- **Nominal to numeric data conversion**

In this step nominal features are converted to numeric data. Suppose F is nominal feature with {fv1, fv2, fv3, fvn} with n distinct nominal values. To convert these nominal values each value fvi is assigned with numeric ID. For example Size is one feature and its set of distinct values is (tiny, very small, small, medium, large, very large, and huge). This set of values is mapped to numeric values by assigning ID as (1, 2, 3, 4, 5, 6, and 7). In this example we have mapped values sequentially but Id can be assigned randomly and Map file generated explicitly.
Output of this step will be database/dataset with all numeric values and Map file which holds the information which numeric ID is for which nominal value.

- **Enhance feature selection for Supervised learning dataset**

In this step co relation of feature Fi is calculated only with class feature using equation 1.In RRW this step is carried out for each feature with all other features i.e. if there are n features then this step is executed $2^{n-1}-1$ times . For example n are 5 then 15 times this equation 1 will be executed. If we consider supervised learning data then importance of the feature is depends solely on its co-relation with class feature therefore whenever there is supervised learning data then there is no need to find co –relation of feature with all other feature except class feature. Therefore equation 1 will be executed only n times if there are n features. For example there are 5 features then equation will be executed 5 times only. This will be very efficient when data is high dimensional when n is like greater than 100.
Input: All features F = {f1, f2, f3….fn}, Class feature, MI mutual information equation, threshold, empty vector V
Process:

```
    For each fi
      MIi = MI (fi,C)
      If MIi < threshold
        Then add fi to Vector C
      End if
```

End For
Select fi randomly from V for feature selection
Output: Selected feature Fi
Watermark encoding, watermark detection and data recovery steps will be same as RRW as discussed in above section.

2) Post-processing
To recover nominal values of originally nominal features, map file is used and numeric ID in the dataset is replaced by respective nominal values.

4. Mathematical Model

Let, S be the system having Input, Processes and Output. It can be represented as,
S = {I, P, O}
Where, I is a set of all inputs given to the System, O is a set of all outputs given by the System, P is a set of all processes in the System.
I = {I1, I2, I3, I4}
I1= D is the supervised learning dataset with
1) Feature set F = {f1, f2, f3…fn} with n number of features and C is class variable or feature
2) With R tuples where each tuple is set of values of all n features and its corresponding class
I2 = Feature Set F = {f1, f2, f3…fn} with n number of features
I3 = Mutual Information threshold
I4 = String to be watermark
P = {P1, P2, P3, P4, P5, P6, P7, P8}
P1 – Process of converting nominal features to numeric features, output of this process will be O1 and Map file O8
P2- Mutual information of each feature is calculated using following formula

$$MI(A, B) = \sum_a \sum_b P_{AB}(a,b) \log \frac{P_{AB}(a,b)}{P_{AB}(a)P_B(b)} \quad (1)$$

Where A is the feature and B is class variable
$P_B(b)$ =marginal probability distribution
$P_{AB}(a,b)$ =joint probability distribution
Output of this process will be O2.
P3 – Finds the features whose values are less than I1 and output will be O3
P4 - Watermark creation using Genetic algorithm, output will be optimized watermark O 4 and O5 set of beta values for each value of selected feature
P5 – watermark embedding in selected feature
For each watermark bit w and for each value r of selected feature
If Brw ==0
Then Dw = Dr + b
V = V U C
Else
Then Dw = Dr –b
V = V U C
Where V is change vector and C is change in the Dr
Output will be O6 and O7
P6 – Watermark decoding
If PC <= 0
Then dtW =1
Else if 0< PC <=1
dtW = 0
Where PC is percentage change
And dtW is detected Watermark bit string
Output will be O7
P7 – Data recovery
O5, O6 and O8 will be used for data recovery
If dtW bit b for r==1
Then Dr = Dwr +B
Else Dr= Dwr –B
Output will be O1
P8 – Post processing
Numeric attributes are converted to its Nominal format using O9 and output will be I1
O= {O1, O2, O3, O4, O5, O6, O7, O8, O9}
O1 = Dataset with all numeric features
O2 - Set of MI value of each feature
O3- Selected feature for watermark encoding
O4 – Optimized watermark data
O5 - Set of beta values for each value of selected feature
O6 – Watermarked dataset
O7 – Change vector
O8- Detected watermark
O9 – Map file from process P1

5. Experimental Results

The goal of the experimental evaluation is to check the results of applying enhanced robust and reversible watermarking for supervised learning data and also to check the time and memory requirements. Experiments will be conducted on core 2 duo with CPU having windows 7, 160 GB hard disk and 2 GB RAM. For database operations MYSQL and MYSQL YOG will be used. Any relational database containing table with nominal attributes can be used for experimental purpose. It is expected that, the proposed method will perform the operations in less time as it only takes n times to calculate MI value of features where existing system requires 2n time where n is the number of features and time is in milliseconds. Proposes system is also able to work on nominal values.
X- axis- Number of Features
Y- axis- Time required in Ms.

Table 1: Existing System vs. Proposed System

Number of features	Existing system (time require ms)	Proposed System (time require in ms)
5	32	4
7	128	6
9	512	8

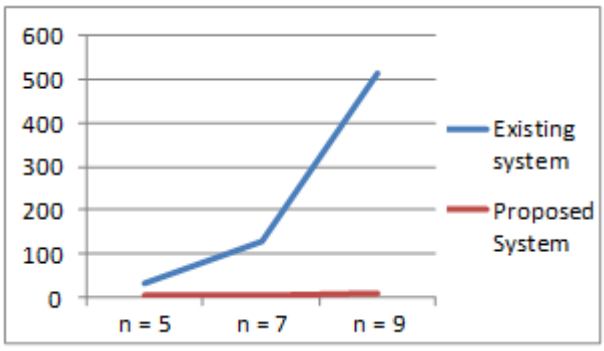

Figure 3: Existing System vs. proposed System

6. Conclusion

This paper describes Watermarking, reversible watermarking And RRW for quality of data and data protection this paper Also describes literature survey, existing system and advantage, disadvantage of given system. In feature selection phase, existing system calculates co-relation of feature with all other features present in the database to find the importance of the feature but when we consider database / dataset of supervised learning training data there is no need to consider all other features. RRW is not applicable to non-numeric features. We introduced a technique which does not require use of all the features instead we pruned the number of features and increase the speed and accuracy of the process. Our method also works on non-numeric data and gives better results, reduces computation, storage and time requirements.

7. Acknowledgement

It is been rightly said that we are built on the shoulder of others. For everything I achieved, the credit goes to all those who had really helped me to complete this work successfully. I am extremely thankful to my Project Guide Prof. R.A.Kudale for guidance and review of this paper work. I would also like to thank the all faculty members of SKNCOE, Pune and my friends who helped me for this work.

References

[1] G. Gupta and J. Pieprzyk, "Database relation watermarking resilient against secondary watermarking attacks," in Information Systems and Security New York, NY, USA: Springer, 2009, pp. 222–236
[2] E. Sonnleitner, "A robust watermarking approach for large databases," in Proc. IEEE First AESS Eur. Conf. Satellite Telecommun., 2012
[3] http://link.springer.com/chapter/10.1007/978-3-642-10772-6_17
[4] E. Sonnleitner, "A robust watermarking approach for large databases," in Proc. IEEE First AESS Eur. Conf. Satellite Telecommun.,2012, pp. 1–6.
[5] K. Jawad and A. Khan, "Genetic algorithm and difference expansion based reversible watermarking for relational databases," J. Syst. Softw., vol. 86, no. 11, pp. 2742–2753, 2013.
http://www.sciencedirect.com/science/article/pii/S0164121213001428
[6] Saman Iftikhar, M. Kamran, and Zahid Anwar, RRW—A Robust and Reversible Watermarking technique for relational data Ieee Transactions on Knowledge and Data Engineering, Vol. 27, No. 4, April 2015