SECURING MEDICAL DATA USING ROBUST AND REVERSIBLE WATERMARKING TECHNIQUE

[1]Ms. K. Devipriya, [2]Ms. M. Lakshmi, [2]Ms. V. Manisha, [2]Ms. N. Manju Bharathi
[1]Assistant Professor, [2]B.E student,
Department of Computer Science and Engineering,
Sri Krishna College of Engineering and Technology, Coimbatore

Abstract-Database Security using RRW deals with securing the numerical medical datasets using watermarking techniques. The Medical databases in hospitals contain important numerical data to be protected against malicious attacks and other data discrepancies. In this project, a technique known as RRW (Robust and Reversible Watermarking) is implemented for securing the data. The Reversible watermarking technique provides protection of ownership rights, data tempering and data integrity, whereas irreversible watermarking schemes only protect ownership rights. A particular attribute is taken into account based on its role by using knowledge discovery. RRW is also evaluated through attack analysis where the watermark is detected with maximum decoding accuracy in different scenarios. The embedded watermark can subsequently be used for proving and claiming ownership of numerical medical datasets.

Index Terms—Reversible watermarking, genetic algorithm, data recovery, data quality, robustness, numerical data.

I INTRODUCTION

Data is very largely used in the today world, increasing the usage of Internet and cloud computing. In different digital formats data can be stored and they are audio, video, images, natural language texts and relational data. Share the relational data by the owners with the virtual storage locations. If the data is make openly available, it will be useful for decision making and knowledge extraction and those datasets are attractive target for attacks. For example document incident attack. According to a survey related to the security, while sharing the data do not consider security and privacy issues in the organizations. The data of health care and their domain can be increased. To ensure the security of ownership protection and tampering proof, watermarking technique is established. The technique can ensure the data recovery along with the ownership protection is that Reversible watermarking technique. The other techniques for ownership protection such as fingerprint, serial codes and data hashing. Transactional watermarking technique is used to monitor and identify the digital watermarking and watermark all the copies with different watermark. Similarly the other techniques can be established with their features and it will have some drawbacks in the reversible watermarking technique. To overcome their drawback we use robust and reversible watermarking technique.

1.1 TECHNIQUES

The commonly known watermarking technique is that Digital Watermarking. To watermark relational databases is different from the process of multimedia watermarking because the fundamental difference in the properties of particular data. The drawback in the digital watermark is that they modify any large content of data the result must be a loss of data quality.

Reversible Watermarking tries to overcome the problem of data quality and this technique keeps the data useful for knowledge discovery. The knowledge discovery is

successful decision making support because high quality of data. In literature the Reversible Watermarking technique is also available but no work can be conducted to overcome the problem that is presence of malicious attack. Attack resilience is the ability to recover the original data and watermark data is the challenging task. To find the most watermark robustness data without significant loss of information. We get a model of optimization bandwidth as a Constraint optimization problem. They optimize a objectives with respect to the variables that are bounded by certain constraints. The ownership rights of those databases are protected from the malicious attack, the data quality may be protected with their constraints. The computational intelligence techniques such as particle swarm optimization(PSO) and Genetic algorithm(GA).

II RELATED WORKS

The first irreversible watermarking technique for relational databases was proposed by Agrawal and Kiernan. Similarly, the first reversible watermarking scheme for relational databases was proposed. In this technique, histogram expansion is used for reversible watermarking of relational database. Zhang et al. proposed a method of distribution of error between two evenly distributed variables and selected some initial nonzero digits of errors to form histograms. Histogram expansion technique is used to reversibly watermark the selected nonzero initial digits of errors. This technique is keeps track of overhead information to authenticate data quality. However, this technique is not robust against heavy attacks (attacks that may target large number of tuples).

Difference expansion watermarking techniques (DEW) exploit methods of arithmetic operations on numeric features and perform transformations. The water-mark information is normally embedded in the LSB of features of relational databases to minimize distortions. Whereas, in RRW, a GA based optimum value is embedded in the selected feature of the dataset with the objective of preserving the data quality while minimizing the data distortions as a result of watermark embedding. Another reversible watermarking technique proposed in [26] is based on difference expansion and support vector regression (SVR) prediction to protect the database from being tampered. The intention behind the design of these techniques is to provide ownership proof. Such techniques are vulnerable to modification attacks as any change in the expanded value will fail to detect watermark information and the original data.

Genetic algorithm based on difference expansion water-marking (GADEW) technique is used in a proposed robust and reversible solution for relational databases. GADEW improves upon the drawbacks mentioned above by minimizing distortions in the data, increasing watermark capacity and lowering false positive rate. To this end, a GA is employed to increase watermark capacity and minimize introduced distortion. This is because the watermark capacity increases with the increase in number of features and the GA runs on more features to search the optimum one for watermarking. However, watermark capacity decreases with the increase in watermarked tuples. GADEW used the distortion measures (AWD and TWD) to control distortions in the resultant data. In this context, the robustness of GADEW can be compromised when AWD and TWD are given high values.

Prediction-error expansion watermarking techniques (PEEW) like incorporate a predictor as apposed 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 water-mark information is embedded in the fractional part of numeric features only. In this particular scenario, the scheme works because the intention of the attacker is to pre-serve the usefulness of the data; otherwise, he can easily compromise the fractional part. RRW is robust, as the water-mark is embedded in the values of numeric features, to make the scheme resilient against such attacks.

III RRW ARCHITECTURE

This section discusses RRW for reversible watermarking of relational databases that improves data recovery ratio. The main architecture of RRW is presented in Fig. 1. RRW includes the following four major phases: (1) Preprocessing (2) Mutual Information (3) Feature selection (4) watermark encoding;(5) watermark decoding and data recovery.

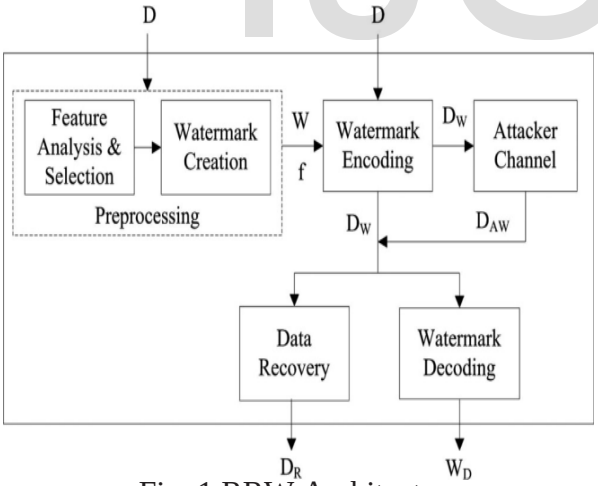


Fig. 1 RRW Architecture

The watermark preprocessing phase computes different parameters for calculation of an optimal watermark. These parameters are used for watermark encoding and decoding. The main focus of watermark encoding phase is to embed watermark information in such a way that it does not affect the data quality. During watermark embedding, data gets modified according to the available bandwidth (or capacity) of the watermark information. The bandwidth of the watermark should be sufficiently large to ensure robustness but not so large that it destroys the data quality. The data owner decides the amount of data modification such that the quality is not compromised for a particular database application before-hand and therefore defines usability constraints to introduce tolerable distortion into the data.

After watermarking, the data is released to the intended recipients over a communication channel that is assumed to be insecure and termed as the “attacker channel” in this research domain. The data may undergo several malicious attacks in the attacker channel. The efficiency and effectiveness of RRW is described through robustness analysis determined by its response to subset insertion, alteration and deletion attacks. The Watermark decoding phase recovers watermark information effectively for detection of the embedded watermark. Data recovery phase mainly comprises the important task of successful recovery of the original data. For a quick reference. In subsequent sections, different phases of RRW are discussed.

3.1 PREPROCESSING

Datasets involves large number of features or large number of tuples. Those data's are contains some noises and symbols. The data set is a comma separated value (CSV) file. The mutual information and the further process cannot be done without finishing the pre process step. In this, the comma delimiter is used to split every feature values and it stored in the database. Important tasks are accomplished

- Selection of a suitable feature for watermark embedding.

3.1.1 FEATURE ANALYSIS AND DESIGN

All the features are ranked according to their importance in information extraction, subject to their mutual dependence on other features. For this purpose, mutual information (MI), is exploited, that is an important statistical measure for computation of mutual dependence of two random variables.

The data owner can define a secret threshold based on MI of all the features in the database. The feature(s) having MI lower than that threshold can be selected for watermarking. The attacker will not attack the features having large MI as in that case the usability of the data will be compromised. Therefore, attacker will be forced to attack the feature(s) with lower MI without concrete knowledge (due to the use of secret threshold) of which features have been watermarked.

3.2 MUTUALINFORMATION

Mutual Information is a well known information theory (concept), statistically measures the amount of information that one feature contains about the other features in a database. In RRW, mutual information is used to select a suitable (candidate) feature from the database for watermarking. the mutual information measure for determining relative importance of features.

Mutual information of every feature with all other features is calculated by using Equation (1).

$$MI(A,B) = \sum_A \sum_B PAB(a,b) log \frac{PAB(a,b)}{PA(a)PB(b)}$$

Where MI(A,B) measures the degree of correlation of features by measuring the marginal probability distributions as PA(a), PB(b) and the joint probability distribution PAB(a, b).

3.3 FEATURE SELECTION

The value of MI of each feature is then used to rank the features. The attacker can try and predict the feature with the lowest MI in an attempt to guess which feature has been watermarked. To deceive the attacker for this particular scenario, a secret threshold can be used for selecting the feature for watermark embedding. In this context, the data owner can define a secret threshold based on MI of all the features in the database. The feature(s) having MI lower than that threshold can be selected for watermarking. The attacker will not attack the features having large MI as in that case the usability of the data will be compromised. Therefore, he will be forced to attack the feature(s) with lower MI without concrete knowledge (due to the use of secret threshold) of which features have been watermarked.

3.4 WATERMARK ENCODING

For the creation of optimal watermark information, that needs to be embedded in the original data, we use an evolutionary technique; GA. GA is a population-based computational model, basically inspired from genetic evolution GA evolves a potential solution to an optimization problem by searching the possible solution space. In the search of optimal solution, the GA follows an iterative mechanism to evolve a population of chromosomes. The GA preserves essential information through the application of basic genetic operations to these chromosomes that include: selection, crossover, mutation and replacement. The GA evaluates the quality of each candidate chromosome by employing a fitness function. The evolutionary mechanism of the GA continues through a number of generations, until some termination criteria is met. Constrained optimized fitness function. 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 information calculation is formulated as a CO problem to meet the data quality constraint of the data owner. A GA is used to create optimal watermark information that includes: (1) Optimal chromosomal string (watermark string of length *l*); and (2) β value. β is a parameter that is computed using GA and represents a tolerable amount of change to embed in the feature values. Once the optimum value of b for each candidate feature A is found, it is saved for use during watermark encoding and decoding. A watermark (bit string) of length l and an optimum value b is used to manipulate the data provided it satisfies the usability constraints .The value β is added into every tuples of the selected feature A when a given bit is 0; otherwise, its value is subtracted from the value of the feature.

Algorithm 1 Watermark Encoding
Input: D, ω, β
Output: Dw, Δ
for ω=1 to *l* do
//loop will iterate for all watermark bits w from 1 to length l of the watermark
for r = 1 to R do
//loop will iterate for all tuples of the data
if br,w= = 0 then
// the case when the watermark bit is 0
changes are calculated and data is watermarked, insert ŋr into r
end if
if br,w= =1 then
// the case when the watermark bit is 1
changes are calculated and data is watermarked, insert ŋr into r
end if
end for
end for
return Dw,r.

3.5 WATERMARK DECODING AND DATA RECOVERY

In the watermark decoding process, the first step is to locate the features which have been marked. The process of optimization through GA is not required during this phase. We use a watermark decoder z, which calculates the amount of change in the value of a feature that does not affect its data quality. The watermark decoder decodes the watermark by working with one bit at a time.

Algorithm 2. Watermark Decoding
Input: Dw, Δ, *l*
Output: WD
for r = 1 to R do
//loop will iterate for all tuples of the data
for b = l to 1 do
//loop will iterate for all watermark bits *b* from 1 to length *l* of the watermark
ŋdr← Dw(r) * ζ
ŋΔr← ŋdr - ŋr
if ŋΔr ≤ 0 then
detected watermark bit (dtW) is 1
else if ŋΔr > 0 and ŋΔr ≤ 1 then
detected watermark bit (dtW) is 0
end if
end for
end for
WD ← mode (dtW(1, 2,….,*l*))
return WD.

DATA RECOVERY

After detecting the watermark string, some post processing steps are carried out for error correction and data recovery. The optimized value of b computed through the GA is used for regeneration of original data.

Algorithm 3. Data Recovery
Input: DW,b
Output: Dr
for r = 1 to R do
//loop will iterate for all tuples of the data

```
    for b = l to 1 do
       //loop will iterate for all watermark bits
b
        from 1 to length l of the watermark
        if dtW(r,b) = = 1 then
            // 0 or 1 watermark bit is detected
            from every tuple r data is recovered
        else
            data is recovered
        end if
    end for
  end for
return Dr.
```

IV RESULTS AND DISCUSSION

Experiments are conducted on Intel(R) Core(TM) 2 Duo with CPU of2.20 GHz and RAM of 3GB. For brevity, heart disease medical dataset, [35] containing more than 300 tuples is selected. RRW was evaluated for: (1) investigating effect on the data quality of the underlying data; (2) robustness against malicious attacks; and (3) restoration of the original data. The data recovery, watermark detection accuracy and effect of RRW on data quality are evaluated using the case study of a heart disease medical dataset. A small set of tuples from the same dataset are also used as an example to illustrate the entire procedure step by step.

Robustness of RRW is demonstrated through an extensive attack analysis. Our results have shown 100 percent accuracy in both watermark detection and data recovery.

age	sex	cp	Trestbps	chol	fbs	Restecg
63	1	1	145	233	1	2
67	1	4	160	286	0	2
37	1	3	130	250	0	0
41	0	2	130	204	0	2
56	1	2	120	236	0	0
62	0	4	140	268	0	2
57	0	4	120	354	0	0
63	1	4	130	254	0	2
53	1	4	140	203	1	2
57	1	4	140	192	0	0
56	0	2	140	294	0	2
56	1	3	130	256	1	2
44	1	2	120	263	0	0
52	1	3	172	199	1	0
48	1	2	110	229	0	0

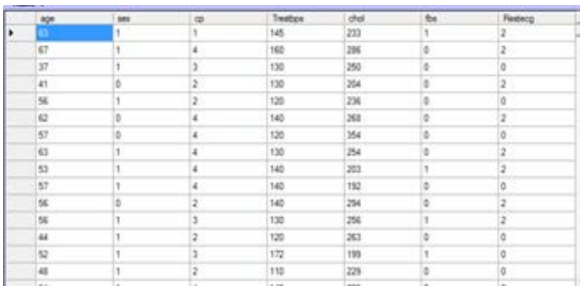
TABLE 1 Original data

Sl_No	Feature	MI
1	age	0.124392796017
2	sex	0.125381927960...
3	cp	0.582478179892...
4	Trestbps	2.428635771585...
5	chol	4.549538034546...
6	fbs	0.027389256107...
7	Restecg	0.162595040718...
8	Thalach	2.759071930273...
9	Exang	0.060256363437...
10	Oldpeak	0.191724792754...
11	Slope	0.295195315828...
12	ca	0.122338677281...
13	Thal	0.867326443413...

TABLE II Mutual Information Values

age	sex	cp	Trestbps	chol	fbs
61.72	-0.28	1	145	233	-0.2
65.72	-0.28	4	160	286	-1.2
35.72	-0.28	3	130	250	-1.2
39.72	-1.28	2	130	204	-1.2
54.72	-0.28	2	120	236	-1.2
60.72	-1.28	4	140	268	-1.2
55.72	-1.28	4	120	354	-1.2
61.72	-0.28	4	130	254	-1.2
51.72	-0.28	4	140	203	-0.2
55.72	-0.28	4	140	192	-1.2
54.72	-1.28	2	140	294	-1.2
54.72	-0.28	3	130	256	-0.2

TABLE III Watermarked Data



age	sex	cp	Trestbps	chol	fbs	Restecg
63	1	1	145	233	1	2
67	1	4	160	286	0	2
37	1	3	130	250	0	0
41	0	2	130	204	0	2
56	1	2	120	236	0	0
62	0	4	140	268	0	2
57	0	4	120	354	0	0
63	1	4	130	254	0	2
53	1	4	140	203	1	2
57	1	4	140	192	0	0
56	0	2	140	294	0	2
56	1	3	130	256	1	2
44	1	2	120	263	0	0
52	1	3	172	199	1	0
48	1	2	110	229	0	0

Table IV After Decoding and Original values

V CONCLUSION

Irreversible watermarking techniques make changes in the data to such an extent that data quality gets compromised. Reversible watermarking techniques are used to cater to such scenarios because they are able to recover original data from watermarked data and ensure data quality to some extent. However, these techniques are not robust against malicious attacks—particularly those techniques that target some selected tuples for watermarking. In RRW a novel robust and reversible technique for watermarking numerical data of relational databases is presented. The main contribution of this work is that it allows recovery of a large portion of the data even after being subjected to malicious attacks. RRW is also evaluated through attack analysis where the watermark is detected with maximum decoding accuracy in different scenarios. A number of experiments have been conducted with different number of tuples attacked.

VI FUTURE ENHANCEMENTS

Our future concern is to watermark shared databases in distributed environments where different members share their data in various proportions. We also plan to extend RRW for non-numeric data stores.

REFERENCES

[1] Y.-C. Liu, Y.-T. Ma, H.-S. Zhang, D.-Y. Li, and G.-S. Chen, “A method for trust management in cloud computing: Data coloring by cloud watermarking,” Int. J. Autom. Comput., vol. 8, no. 3, pp. 280–285, 2011.

[2] P. W. Wong and N. Memon, “Secret and public key image watermarking schemes for image authentication and ownership verification,” IEEE Trans. Image Process., vol. 10, no. 10, pp. 1593–1601, Oct. 2001.

[3]P. E. Gill, W. Murray, and M. A. Saunders, “Snopt: An sqp algorithm for large-scale constrained optimization,” SIAM Rev.,vol. 47, no. 1, pp. 99–131, 2005.

[4] 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.

[5] J.-N. Chang and H.-C. Wu, “Reversible fragile database watermarking technology using difference expansion based on SVR prediction,” in Proc. IEEE Int. Symp. Comput., Consum. Control, 2012, pp. 690–693.

[6] 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.

[7] M. E. Farfoura and S.-J. Horng, “A novel blind reversible method for watermarking relational databases,” in Proc. IEEE Int. Symp. Parallel Distrib. Process. Appl., 2010, pp. 563–569.

[8] D. M. Thodi and J. J. Rodriguez, "Prediction-error based reversible watermarking," in Proc. IEEE Int. Conf. Image Process. 2004, vol. 3,pp. 1549–1552.

[9] D. M. Thodi and J. J. Rodriguez, "Reversible watermarking by prediction-error expansion," in Proc. 6th IEEE Southwest Symp.Image Anal. Interpretation, 2004, pp. 21–25.

[10] D. M. Thodi and J. J. Rodriguez, "Expansion embedding techniques for reversible watermarking," IEEE Trans. Image Process., vol. 16, no. 3, pp. 721–730, Feb. 2007.

[11] M. E. Farfoura, S.-J. Horng, J.-L. Lai, R.-S. Run, R.-J. Chen, and M. K. Khan, "A blind reversible method for watermarking relational databases based on a time-stamping protocol," Expert Syst. Appl., vol. 39, no. 3, pp. 3185–3196, 2012.