ISSN:0975-9646

Watermarking of Dataset Using Usability Constraints Model

Joycee .A.A,
M.E. Computer Science and Engineering,
Sathyabama University,
Chennai, India.

Pio Sajin,
Assistant Professor,Department of computer science and engineering,
Sathyabama University,
Chennai, India.

***Abstract*- In the emerging field of "Sharing Dataset" with the deliberate recipients, protecting ownership on the datasets is becoming a challenge in itself. Watermarking is the commonly used mechanism to enforce and prove ownership for the digital data in different formats like text, database, software, image, video, audio, image and relational database. Major issue is that cannot protect the contained information in the datasets. The most important challenge in watermarking data mining dataset is: how to preserve knowledge in features or attributes during the embedding of watermark bits? To meet this requirement, embedding watermark in datasets using usability constraints. Owner defines a "usability constraints" to provide distortion band within which the values of a feature can change for each feature. In addition to this, the inserted watermark should be imperceptible and robust against any type of attacks.**

***Index Terms*- watermarking datasets, data usability, right protection, knowledge-preserving, ownership preserving.**

I. INTRODUCTION

The datasets generate from large databases are mined to extract hidden information and patterns that are useful for decision makers to make effective and efficient. Knowledge –driven data mining expert systems cannot be design and develop until the owner of data is willing to outsource the dataset with corporations. Organizations share their datasets and the business challenge to data mining experts to find novel solutions to the posted problem [1].

A watermark can be applied to any database relation having attributes. To preserve the knowledge in the dataset, we need to ensure the predictive ability of a feature or an attribute is preserved; as a result, the classification accuracy of the dataset remains intact. The process of defining "Usability constraints" is dependent on the dataset and its intended application. To best of our knowledge, no technique has been proposed to model the usability constraints for watermarking of dataset.

In this paper, we propose a model for identifying the "usability constraints" which must be enforced while embedding watermark in a dataset. It uses three different optimizers to find an optimum watermark. The main contributions of our paper are:

- In proposed technique, we define usability constraints on a dataset, which not only preserves the knowledge contained in the dataset but also ensures the robustness of the inserted watermark.
- Integrated our model in new knowledge preserving watermarking scheme to validate its effectiveness and efficacy.

The paper is designed as follows: Section II describes the related work, section III describes the existing system, section IV describes the proposed system, section V describes watermarking scheme, section VI conclusion.

II. RELATED WORK

In the work of R Agrawal, J Kiernan [6], the first technique proposed for watermarking numeric attributes in a database. In this technique, MAC (Message Authentication Code) is calculated with the help of a secret key to identify the candidate tuples. J Kiernan, P Haas, R Agrawal [3] proposed watermarking tuples in a relational database uses signals. It inserts watermark with multiple bits on multiple tuples. But it needs to improve the optimization.

S Krishnaswamy, M Kwon, D Ma, J Palsberg [4], experimented with a watermarking system for java which embeds a watermark in dynamic data structures. They show that watermarking can be done efficiently to gradually increase the size of code, heap space and execution times. M Crogan, V Raskin, M Atallah [5], presented two main results in the area of information hiding in natural language text. Semantically based scheme improves the information hiding capacity through two techniques: i) modifying the granularity of individual sentences ii) halving the number of sentences affected by watermark.

R Sion, S Prabhakar [7], presented a tuples based watermarking technique for relational database but it is not applicable for data mining datasets because they are not preserving the information contained in the dataset. A Ghafoor, E Bertino, M Shehab [8], describes a portioning based watermarking technique. The process of watermark insertion as a constraint optimization problem and tested GA (Genetic Algorithm) and PS [9] (Pattern Search) optimizers. After testing they select PS because it is able to optimize in real time.

M Farooq, M Kamran [10], recently proposed a technique protecting ownership of EMR (Electronic Medical Records) system. In this information gain is used to find the predictive ability of all features in the EMR. The least

predictive ability of numeric features is selected to embed watermark bits. This technique is only limited for information gain not for other feature selection schemes. This watermarking technique is limited to numeric features only.

The current work is focussed on developing a formal model to define "Usability Constraints" for watermarking of dataset in such a way that the watermark is not only robust but the dataset knowledge also preserved. And we also provide a mechanism to group the dataset based on high ranking features and then watermarked. Because low ranked features are easily hack by an attacker by launching malicious attacks. And we proposed watermarking for numeric and non- numeric attributes.

III. EXISTING SYSTEM

Watermarking techniques enact a vital role in addressing the ownership problem. Such techniques allow an owner of a data to embed imperceptible watermark into the data. The datasets are watermarked and directly send to the client system. In this system, the attacker can easily change or update the data and create some copy of datasets.

Disadvantage Of Existing System

- It does not preserve the knowledge contained in the dataset.

IV. PROPOSED SYSTEM

In this paper, we implement two contributions: i) a model which derives usability constraints for all kinds of datasets. ii) A new watermarking technique works for numeric, non-numeric, strings and image datasets. Our system takes input as a dataset, models the usability constraints during the watermark embedding in the dataset. Watermark embedding technique is used to preserve the watermarked dataset.

The proposed system, logically groups the data into different clusters based on the ranking for defining local usability constraints for each group. Identify the vital characteristics of a dataset which need to be preserved during watermarking. Ensure watermark security by using data grouping and secret parameters.

A Formal Model For "Usability Constraints"

We present our model to define "Usability Constraints". It is used to preserve the data during the process of inserting watermark in the dataset. It provides a distortion band within which the values of a feature can change for each feature. In this paper, three different constraints are used to watermark the dataset. They are:

i) Local usability constraints
ii) Global usability constraints
iii) Image usability constraints

Local Usability Constraints

Local usability constraints L_i, is a tuple initiating mutual information I (M) of the feature M in a particular data group. It can also represent as,

L_i=I (M) (1)

It is used to watermark features in a group and they are applied at a group level only.

Global Usability Constraints

Global usability constraints G is a tuple that consists of features set produce by different feature selection schemes on that dataset. It enforced both at a group level and at the global dataset level. The features set can be applied to a group or a dataset should remain unaltered.

Image Usability Constraints

Image usability constraints I is a tuple that embeds data (watermark) into multimedia object to protect the owner's right to that object. The watermark is embedded by directly modifying the pixel values. The image constraints are more efficient and secure compare with normal textual watermarking.

Definition: Tuple

A tuple $_T$ is an ordered list of elements. The tuple is used as an essential unit for referring different parameters of a dataset.

Definition: Watermark Embedding

Watermark embedding is a transformation of dataset D_O to D_W after embedding a watermark W. It is also represent as,

$(D_O, W) \rightarrow D_W$ (2)

The dataset D_W is shared with a deliberate recipient, therefore; the information lost during the process of watermark embedding to preserve the knowledge.

Architecture For The Proposed System

Architecture diagram shows the relationship among different parts of the system. It is used to clearly understand the whole process.

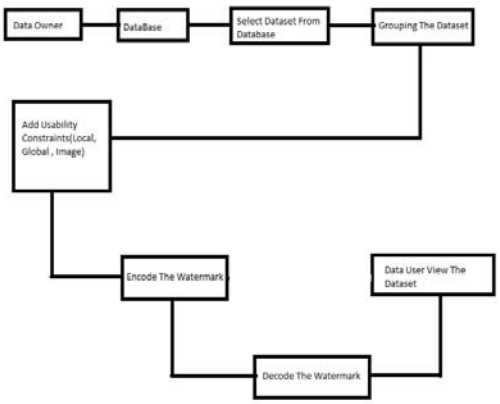

Fig1. *Architecture diagram for the proposed system*

The classification potential features are used to logically group features of the dataset into no overlapping groups [Fig1]. The watermark is optimized and embedded to ensure the usability constraint modelled. The watermark embedding technique is that,

i) Identify the vital characteristics of a dataset which need to preserve during the watermarking.
ii) Ranking the features based on the classification potentials.
iii) Logically grouping the data into different clusters based on this ranking for defining the local usability constraints.
iv) Defining global usability constraints for the complete dataset.
v) Defining image usability constraints for the image dataset.
vi) Modelling the usability constraints so that the learning statistics of classifiers are preserved.

vii) Optimizing the watermark embedding such that all usability constraints can remain intact.
viii) Ensure watermark security by using data grouping and secret parameters.

The classification potential is used to make different groups of a dataset so that the features with high classification potentials are last modified during the process of watermarking. Watermarking process should not modify the original value of a candidate feature.
We can conclude that, the usability constraints, defined the process of watermarking, to ensure knowledge preserving and lossless watermarking.

Advantage of Proposed System

i) It is used to preserve the knowledge contained in the dataset because of using usability constraints.
ii) Data user can view the dataset as the original data after the watermarking process.
iii) They can view the watermarked dataset but can't make any changes.
iv) Knowledge preserving and data lossless.

V. WATERMARKING SCHEME

The watermarking scheme results in Zero information loss. Two main phases in watermarking scheme:

i) Watermark Encoding.
ii) Watermark Decoding.

A. WATERMARK ENCODING

Different steps are involved in watermark encoding phase [Fig 2]. They are,

i) Feature Ranking.
ii) Classification Potential Computation.
iii) Data Grouping.
iv) Refined Usability Constraints.
v) Selecting Data for Watermarking.
vi) Watermark Embedding.

Feature Ranking

- Logically group the data into' n' overlapping partitions.
- Define usability constraints to information loss is zero.
- Ranking is done using information measure.
- Rank all the feature which is present in the dataset and it is stored in a vector.

Classification Potential Computation

It is important to compute the amount of change that a feature can tolerate during the watermarking process. We proved that,

- The features with high classification potential can tolerate only small changes.
- The top ranked features shows zero tolerance towards any change.

Data Grouping

The grouping function is applied on every feature of an input dataset. The groups are logical and it cannot be separated from one another. In earlier work, the data grouping is applied for low ranked features during watermark. So it can be easily attack by an attacker.

We use the groups to define all the usability constraints. Empty group will be omitted during the optimization phase. In the proposed system, the data group can be applied for high ranked features. In the new approach, an attacker cannot easily build an attack by filtering the ranked features.

Refined Usability Constraints

Refine the usability constraints into three types: local usability constraints, global usability constraints, image usability constraints.

- Global constraints applied for the whole dataset.
- Local constraints applied for the logical group of the dataset.
- Image constraints applied for the multimedia object.
- All constraints are applied to the input dataset to watermark.
- Visible and invisible watermarking is done through the usability constraints.
- Local constraints are defined by mutual information.
- Tolerance alteration is maximized for all features in a group.
- Predictive ability of each feature calculated by feature selection scheme and it is preserved.

Select Data For Watermarking

An important step in watermarking of a dataset is to select relevant rows in which the watermark will be inserted. We use a parameter to store information about the selected rows. Main purpose is to insert a watermark for the selected rows to preserve the data presented in the dataset.

Watermark Embedding

Watermark embedding technique is applied for the input dataset based on the feature ranking. The features to be watermarked are,

i) Watermarking non-numeric features.
ii) Watermarking numeric features.

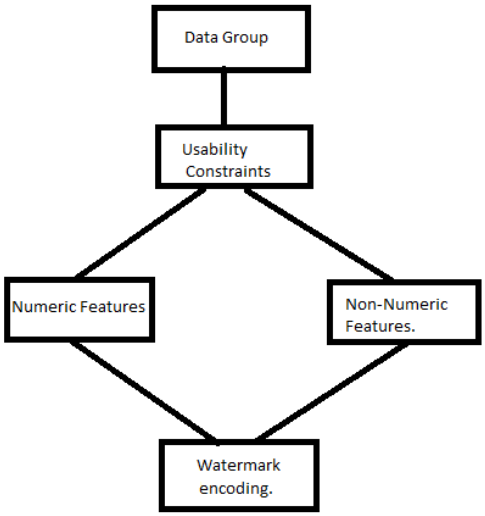

Fig.2 *Watermark Encoding*

According to the above features the data owner encrypts the data using the model of watermark embedding.

Watermarking Non-Numeric Features

Data grouping is not performed for the non- numeric because our watermark embedding technique does not

bring any change in the values of such features. The process is,

- Using sequence of binary bits to embed watermark in a dataset.
- Secret hash value for each row is calculated by pseudo random sequence generator.
- Secret order does not bring any change in the dataset.
- If the row is repeated by the same class label then same hash value will be generated.
- After embedding the final bit, it is stored to use it during the watermark decoding.

Watermarking Numeric Features

The watermarking numeric features are used to maximize the tolerable alternations. The constraints are verified locally for each logical group. The global constraints are verified for the whole dataset. It has the ability to locate the local and global optimum in the search space. The numeric features in a group are marked with bit 1as positive; and with bit 0 as negative.

B. WATERMARK DECODING

Different steps are involved in watermark decoding [Fig 3]. They are,

i) Watermark Decoding From Non-Numeric Features.

ii) Watermark Decoding From Numeric Features.

Watermark Decoding From Non-Numeric Features

The watermark decoding is the reverse process of watermark encoding. The process is defined as,

- Hash value for each row is calculated using pseudo random generator in watermark embedding.
- The watermarked values are stored in the database. It can be view only by the data owner.
- Data user can view the watermarked dataset but they can't change the information.
- During watermark encoding the values which is stored in the database are taken to process watermark decoding.
- It is difficult for data user to find the secret hash values of watermarking which is stored by the data owner.
-

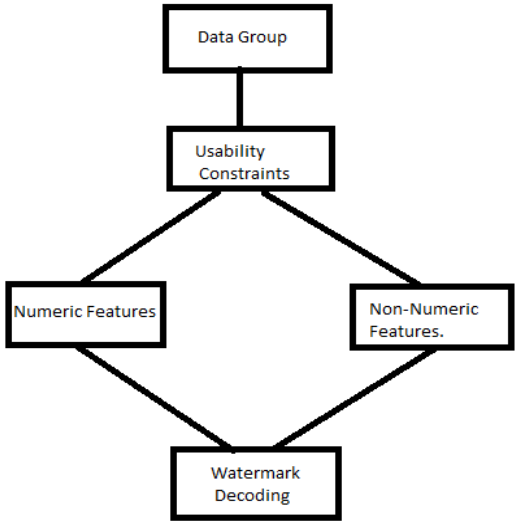

Fig.3 Watermark Decoding

Watermark Decoding From Numeric Features

Watermark decoding from numeric features also compute the same process of watermark encoding of numeric features. Based on the encoding results which is stored in the database are used to decode the watermarked datasets. Without the stored procedure values it is difficult to decode the watermarked dataset. Because of this reason we say that, it is difficult for attacker to decode the knowledge present in the input dataset.

Experimental Design

The experimental diagram shows the total process of a system [Fig 4]. First, it logically group the data based on the feature ranking.

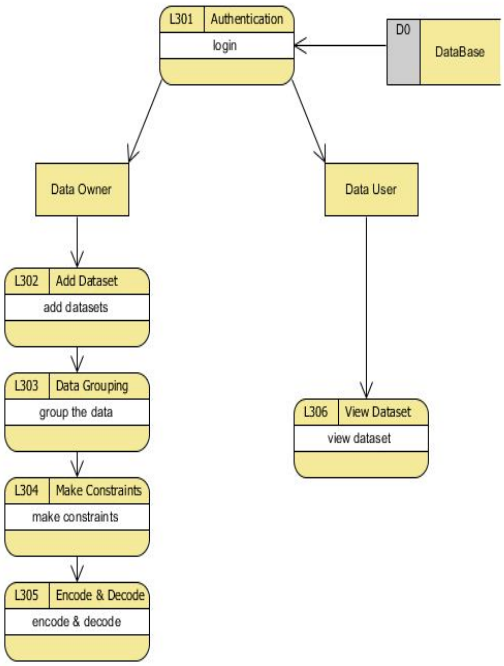

Fig4. Data Flow Diagram

The authorized user can only view the dataset. Even though, the authorized user can also view the dataset not to modify the data. After grouping the data based on the feature selection, then add usability constraints to every group which is present in the input dataset. Finally, watermark embedding is defined in the usability constraints to preserve the information.

Applications

- Television Network: In this application identification, signature of owner we can use watermark.
- TNPSC exam: In this application, easily identify which district question paper released before exam.

Corporate Companies use watermarking techniques for outsourced datasets.

VI. CONCLUSION

In this paper, we proposed a new watermarking scheme to define a usability constraint to preserve the knowledge

contained in the dataset (i.e. data lossless). The benefits of our techniques are:

1) High ranked features are grouped together to apply constraints.
2) Because of using usability constraints we maximize the lossless data.
3) Preserve the knowledge contained in the dataset.
4) Watermark decoding is difficult for the attacker to build attack.
5) Enhanced the watermark technique from numeric features to non-numeric and image features with more watermark security.
6) A new approach "usability constraint" is defined to preserve the dataset.

To my best, no technique in the literature exists that automatically computes "usability constraints" for a dataset to preserve the knowledge contained in the dataset. The proposed system is useful for the customers to share datasets with data-mining experts (corporations) by protecting their ownership. The future work can be extended to video, audio features.

REFERENCES

[1] Kaggle's contests: Crunching Numbers for Fame and Glory 2012[online]. Available: http://www.businessweek.com

[2] Patients Sue Walgreens for Making Money on their data 2012[online]. Available: http://healthcareitnews.com

[3] P Haas, R Agrawal and J Kiernan, "*Watermarking relational data: Framework, algorithms and analysis*", The VLDB Journal, vol.12, no.2, pp. 157-169, 2003.

[4] J Palsberg, S Krishnaswamy, D Ma, M Kwon, Q Shao and Y Zhang,"*Experience with software watermarking*", in Proc. 16th computer security applications conf. 2000, pp. 308-316.

[5] M Atallah, V Raskin, M Crogan, C Hempelmann, F Kerschbaum, D Mohamed and S Naik, "*Natural language watermarking: Design, analysis and proof-of-concept implementation*", in information hiding. New york, NY, USA: Springer, 2001, pp. 185-200.

[6] R Agrawal and J Kiernan," *Watermarking relational databases*", in proc.28th Int. conf. Very Large Databases, 2002, pp. 155-166.

[7] R Sion, M Atallah and S Prabhakar, "*Rights protection for relational data*," IEEE Trans. Knowl. Data Eng., vol.16, no.12, pp. 1509-1525, Dec. 2004.

[8] M Shehab, E Bertino and A Ghafoor, "*Watermarking relational databases using optimization-based techniques*," IEEE Trans.Knowl.Data Eng., vol.20, no.1, pp. 116-129.jan.2008

[9] R Lewis and V Torczon, "*Pattern search methods for linearly constrained minimization*", Institute for Computer Applications in Science and Engineering, NASA Langley Research Center, Hampton, VA, USA, 1998.

[10] M Kamran and M Farooq, "*An information-preserving watermarking scheme for right protection of EMR systems*", IEEE Trans. Knowl. Data Eng., vol.24, no.11, pp. 1950-1962, Nov.2012.

[11] M Pinsker, "*Information and information stability of random variables and processes*", San Francisco, CA, USA: Holden-Day, 1960.

[12] M Kamran and M Farooq," *A formal usability constraints model for watermarking of outsourced data mining datasets*", Tech. Rep. TR-59-Kamran, 2012.