ISSN:0975-9646

Watermarking of Datasets Using Usability Constraints

Anuj Shrivastav
Computer Science and Engineering
SRM University
Kattankulathur, Chennai, Tamil Nadu, India

R.Vidhya
Assistant Professor
Computer Science and Engineering, SRM University
Kattankulathur, Chennai, Tamil Nadu, India

Abstract—**In this world the watermarking is common technology but now days we can save our data and can say that it is our datasets, its called "Proving the ownership".so we can provide the watermark to each datasets. We know that from the large datasets are being mined to extract hidden knowledge and patterns which assist in decision making. So we use the "knowledge-driven", but Data mining activity is not possible without sharing the "datasets "between the data owner and data mining expert. So we introduce the "usability constraint", it is define by the owner for the each type of dataset to preserve the contained knowledge.**

I. INTRODUCTION

In this world many application so that many data will be generate so all data should be safe and should not be taken by the unauthorised user, so we use the watermark for each datasets and proving the ownership. The watermarking technique should be robustness and embedded watermark should be imperceptible. Watermarking should be prove the ownership for the digital data in different formats like audio, video, image, relational database, text and software [3 ,5]. The most important challenge in watermarking Datasets is that "how to preserve knowledge in features or attributes "during the embedding of watermark bits? For the preserve the knowledge in the datasets we have to confirm that the predictive ability of a feature .We know that the owner define the usability constraint so it also provide the distortion band because the owner can be change the features values. The classification accuracy of the dataset should be unaltered. Watermarking techniques have been developed for audio, video, images, and Text data, and also for software, natural language text, relational datasets, numeric and non-numeric value. Many data as for example of weather data, medical data, power Consumption, stock market data, consumer behaviour data and scientific data, we can embed the watermark these all type data's. The all data can be tolerate a small amount of error with respect to their usability constraint. These watermark technique is resilient and robustness.

A "watermark" is a signal that is securely, imperceptibly, and "robustly" embedded into original content such as an image, video, or audio signal, producing a watermarked signal. The watermark describes information that can be used for proof of ownership or tamper proofing. Two type of watermark are 1. Robust and 2. Fragile. *Robust Watermark:* for proof of ownership, copyrights protection (Signature and data are the same object)

Fragile Watermark: for tamper proofing, data integrity. (Integrity information is embedded in the data).The most question is that Why we use the watermarking? Real-world datasets can tolerate a small amount of error without degrading their usability as example is Meteorological data used in building weather prediction models, the wind vector and temperature accuracies in this data are estimated to be within 2.5 m/s and 0.8 ºC. Such constraints bound the amount of change or alteration to that can be performed on the data. Usability constraints are application dependent. Alterations performed by the watermark embedding should be unidentifiable by the human visual system in images/video. For consumer behavior data. Watermarking should preserve periodicity properties of the data. For the right protection we can use the "fingerprinting" but problem is that the owner define the different usability constraints for the same data sets.so it is not possible for it. The major contribution of our paper is that, we define a novel formal model for identifying the essential usability constraints which must be enforced while embedding watermark in a dataset.

1. The proposed technique is independent and numeric or nonnumeric datasets, the "usability constraints" on a dataset it should be preserves the knowledge contained in the dataset and robustness of an insert watermark.
2. We will integrate the new knowledge-preserving watermarking scheme and check the efficacy and effectiveness.
3. The new scheme will compare with the exiting techniques and enhanced the security.
4. In the last we have to do the experiments with much type of data sets and use the machine learning classifier (ML) and get the classification accuracy.

II. RELATED WORK

It is the first technique to define the usability constraint to watermarking data mining. Some description to relate to this watermarking,

1. R.Agrawal [1] define the message authenticated code (MAC),and calculate the MAC of the numeric attributes with the help of a secret key to identify the candidate tuples.
2. Sion [2] defines the watermarking technique, it based on marker tuples. It is only for relational databases.
3. Shehab [3] partitioning based database technique, it define 2main point that are: *Genetic algorithm (GA) Pattern search (PS)* (only for Real-Time) [4] Optimizers. These all techniques are good for embedded the watermark but these technique are not defining the "usability constraint". In this technique we got all the parameters as secret key, MAC, GA, PS etc.
4. M. Kamran and Muddassar Farooq [5] define the usability constraint with help of Electronic medical records (EMR) system.it based on predictive ability.

A multimedia object consists of a large number of bits, with considerable redundancy. Thus, the large watermarks hiding bandwidth. The relative spatial/temporal positioning of various pieces of a multimedia object typically does not change. Tuples of a relation on the other hand constitute a set and there is no implied ordering between them. Portions of a multimedia object cannot be dropped or replaced arbitrarily without causing perceptual changes in the object. However, a pirate of a relation can simply drop some tuples or substitute them with tuples from other relations. The numeric feature(s) with the least predictive ability are selected to embed watermark bits to confirm the information-preserving characteristic. Some important characteristics of dataset that play a important role in classification of the dataset. The major contribution of the technique is that the information-preserving watermarking, it does not describe any mechanism to formal novel model the "usability constraints". This is only for the selected numeric features only. We start the comparison, so we have to focus on developing a formal novel model to describe the "usability constraints" for watermarking of data mining datasets .in the other way can be say that the watermark is robustness and preserve the knowledge contained in the datasets. For more, explanation we define the groups a technique to logically group the dataset into groups, it describe the most important (high ranked) attributes or feature also be watermarked during watermarking. This is the vital enhancement because if the attacker can be attack on the low ranked features which they are watermarked with any compromising the quality of the data and should be great extent.in this model we have to do the grouping for the data, the data owner to embed a watermark in high ranked attributes or features and preserving the knowledge .Attacker has access to only the watermarked data set. The attacker's goal is to weaken or even erase the embedded watermark and at the same time keep the data usable. It is called the "Attacker's Dilemma". Attacker has the 3 type possible attack, 1) Tuple deletion 2) Tuple alteration 3) Tuple insertion. Now we can embed the watermark in any type of feature as numeric and non-numeric [6].The watermark should be invisible because the local constraint can be visible and global constraint should be invisible. The data owner will encode all the data and valid user could be take the permission form the data owner .the data owner check the user and give the useful data and it will decode the data for the user.

III. APPROACH

We present two contributions: (1) a novel framework Model which derives usability constraints for all kinds of datasets; and (2) a new watermarking technique that works for numeric, nonnumeric and strings datasets. Our system takes the dataset as an input, models the "usability constraints" to be enforced during the watermark embedding in the dataset. Later it uses three different optimizers to find an optimum watermark that meets the relative constraints. The predictive ability of features, present in the dataset, is calculated and the features are ranked on the basis of computed predictive ability. Using these ranks, the next step is to generate the logical groups of features. In this step, "local usability constraints" are defined for each logical group. Similarly, the "global usability constraints" are also defined that are applicable for the whole dataset. Finally, both types of constraints are used to build a meta-constraints model that is given as an input to the watermarking scheme.

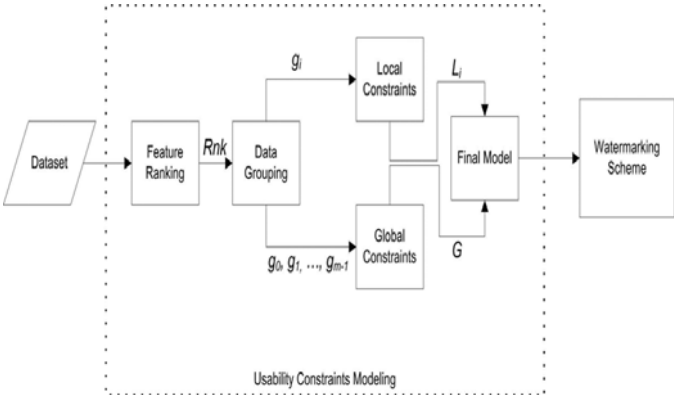

Figure. 1

IV. EXISTING TECHNIQUE:

In Fig.1 earlier existing systems the datasets will be watermarked and directly send to the client system, in these systems while sending datasets from server to client attacker easily can change or update the data and create same copy of datasets. These all technique is good for embedded the watermark but these technique is not define the "usability constraint".

Existing Technique explanation:

Resilient watermarking techniques

In this technique embedding water marking data set without using usability constraint. It supports only numeric features. We have presented a resilient watermarking technique for relational data that embeds watermark bits in the data statistics. The watermarking problem was formulated as a constrained optimization problem that maximizes or minimizes a hiding function based on the bit to be embedded.

Genetic algorithm and pattern search techniques were employed to solve the proposed optimization problem and to handle the constraints.

Some Drawbacks:

1. Watermarking of data mining datasets does not preserve the knowledge contained in the dataset.
2. It not concentrates on usability constraints.
3. Only focus on numeric future.
4. Deleting and changing water marking is difficult.
5. Notable group data.

According to Agrawal:

Agrawal technique based on tuples.We have to find the all tuples and in the tuples also find the numeric data, so it is very long method but it define the watermarking can be possible in all data sets.

1. Watermarking of numerical data.
2. Technique dependent on a secret key.
3. Uses markers to locate tuples to hide watermark bits.
4. Hides watermark bits in the least significant bits.

Weaknesses:

1. No provision of multi-bit watermark, all operations are dependent only on the secret key.
2. Not resilient to alteration attacks. Least Significant Bit (LSB) can be easily manipulated by simple numerical alterations Shift LSB bits to the right/left.
3. Requires the presence of a primary key in the watermarked relation.
4. Does not handle other usability constraints such as: Category preserving usability constraints.

According to sion:

Sion describe is almost same from the Agrawal but it define the primary key and give the concept of normalization is possible in data set when we embed the watermark. This technique not for the data mining because it not preserve the knowledge contain in the data sets.

1. Watermarking of numerical data.
2. Technique dependent on a secret key.
3. Instead of primary key uses the most significant bits of the normalized data set.
4. Divides the data set into partitions using markers.
5. Varies the partition statistics to hide watermark bits.

Weaknesses:

1. Watermark suffers badly from watermark synchronization error cause by
   a. Tuple deletion attacks.
   b. Tuple addition attacks.
2. No optimality criteria when choosing the decoding thresholds
   a. Errors even in absence of attacker.
3. No clear systematic approach for manipulating data
   a. Only a very small space of the feasible data manipulations investigated.

COMPARISION:

Existing system	***Proposed system***
It only concentrate on numeric feature	It concentrate both numeric feature and non numeric feature
Resilient watermarking technique	Watermark Embedding Technique
It not concentrate on usability constraint	It concentrate on formal model usability Constraint
Delete and changing watermark is difficult	Delete and changing watermark is easily
No ranked feature.	Ranked feature used for watermark embedding.

V. PROPOSED ARCHITECTURE

Ownership rights on outsourced relational database are very crucial issue in today's internet environment and in many content distribution applications, because the rapid growth of the internet and related technologies offered an unprecedented ability to access and redistribute digital content. In this paper, Fig.2 we present two contributions such as a novel framework model which derives usability constraints for all kinds of datasets; and a new watermarking technique that works for numeric, nonnumeric and strings datasets. Our system takes the dataset as an input, models the "usability constraints" to be enforced during the watermark embedding in the dataset. Later it uses three different optimizers to find an optimum watermark that meets the relative constraints.

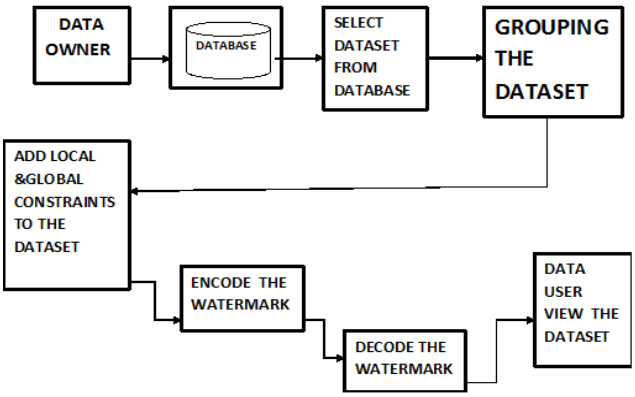

Figure.2

In this project embedding watermark in dataset using usability constraint, in exists technique cannot use usability constraint, first of all group the data from dataset then chose local constraint or global constraint and find numeric feature Or non-numeric feature, non-numeric feature mean cannot group data and less alteration numeric feature mean no loss for mutual information. Ownership rights on outsourced relational database are very crucial issue in today's internet environment and in many content distribution applications, because the rapid growth of the internet and related technologies offered an unprecedented ability to access and redistribute digital content. In this paper, we present two contributions such as a novel framework model which derives usability constraints for all kinds of datasets; and a new watermarking technique that works for numeric, nonnumeric and strings datasets. Our system takes the dataset as an input, models the "usability constraints" to be enforced during the watermark embedding in the dataset. Later it uses three different optimizers to find an optimum watermark that meets the relative constraints.

Proposed System Technique Explanation

Watermark Embedding Technique

Step 1: The classification potential of each feature is calculated using mutual information and it is stored in a vector. The threshold is computed using a vector of classification potentials. The classification potential of features and are then used to logically group features of the dataset into no overlapping groups.

Step 2: The watermark is optimized and embedded in this stage while enforcing the usability constraints modelled

Advantage:

1. Logically grouping the data into different groups (clusters) based on this ranking for defining local usability constraints for each group.
2. Ensuring watermark security by using data grouping and secret parameters.
3. Identifying the vital characteristics of a dataset which need to be preserved during watermarking.

Application:

1. Television network: In Television Network based application we can use watermark embedding technique in order to prove the ownership of channel distributor.
2. TNPSC: In TNPSC Examination centre we can detect the question paper leakage before examination easily by embedding watermark to question paper dataset.

VI. CONCLUSIONS

A novel knowledge-preserving and lossless usability constraints model and a new watermarking scheme have been proposed for watermarking data mining datasets. The benefits of our techniques are: identifying the vital characteristics of a dataset which need to be preserved during watermarking; ranking the features on the basis of their classification potentials; logically grouping the data into different groups based on this ranking for defining local usability constraints for each group; defining global usability constraints for the complete dataset; modeling the local and global usability constraints in such a manner so that the learning statistics of a classifiers are preserved; optimizing the watermark embedding such that all usability constraints remain intact; ensuring watermark security by using data grouping and secret parameters.

REFERENCES:

[1] R. Agrawal and J. Kiernan, "Watermarking relational databases," in Proc. 28th Int. Conf. Very Large Data Bases, 2002, pp. 155–166

[2] 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.

[3] 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

[4] 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.

[5] 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.

[6] M. Kamran and M. Farooq, A Formal Usability Constraints Model for Watermarking of Outsourced Data Mining Datasets Tech. Rep. TR-59- Kamran, 2012.