ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 4, Issue 7, July 2015*

# Rule-based Watermarking Model for Numeric and Non-numeric Datasets

**Kanchan S. Rahinj**

Student of ME, Information Technology, Amrutvahini College of Engg, Sangamner, India

**Abstract**: The large datasets are being mined to take out hidden information and patterns that assist decision makers in making effective, well-organized and well-timed decisions in an ever increasing competitive world. This kind of knowledge-driven data mining activity is impossible without sharing the datasets between the owners of datasets and data mining experts or corporations. As a result, protecting ownership by embedding a watermark on the datasets is becoming applicable. The main challenge in watermarking to be mined datasets is, How to conserve knowledge in features or attributes? The owner needs to manually define Usability constraints for every type of dataset to protect the contained knowledge. The main contribution is a novel formal model that facilitates a data owner to describe usability constraints to preserve the knowledge contained in the dataset in an automated manner. The formal model aims at preserving "classification potential" of each one characteristic and other most main characteristics of datasets that participates in an important role during the mining process of data; as an end result, learning statistics and decision-making rules also remain unbroken. I will implementing a model and integrating it with a new watermark embedding algorithm to demonstrate that the inserted watermark not only conserve the information contained in the dataset, but also significantly increases watermark security as compared with existing systems.

**Keywords**: knowledge-preserving and ownership preserving data mining, Data usability, watermarking datasets, right protection.

## I. INTRODUCTION

The large datasets generated from very large databases are being mined to take out secret knowledge and patterns that are showing useful for decision makers to build effective, efficient and timely decisions in this world. This type of data driven data mining expert schemes cannot be intended and developed in anticipation of the owner of data is ready to distribute or outsource that dataset among data mining experts or corporations. This system is develop

1. A novel framework model which develop usability constraints for all kinds of datasets and
2. A novel watermarking technique that facilitates for numeric, non-numerical and strings datasets.

The system receives the dataset as an input, forms the usability constraints to be imposed during the watermark embedding in the dataset. Then it applies three different type of optimizers to find an best possible watermark that gather the comparative constraints. In the emerging field of sharing datasets with the intended receivers, protecting the rights on the datasets is becoming a main challenge in itself. In recent times, an article reported the illegitimate sale of patient data and the concerned patients have charged the original hospital for breaching their privacy. An even larger concern is that the recipient may try to take credit for the contribution towards knowledge discovery and data mining (KDD) by claiming the wrong ownership of the public data. To alleviate these threats, a nondisclosure contract is usually signed with the receiver binding him that he will not share the dataset and will also not argue the rights of the data. If the recipient breaches the agreement, the legitimate data owner can only go to court him if he can verify in a court of law his rights more than the dataset. Watermarking is the usually used mechanism to enforce and confirm ownership for the digital data in different forms as image, audio, relational database, video, text and also software [3][5]. The important challenge in watermarking data mining datasets one is: how to save knowledge in features or an attributes during the insertion of watermark bits?

In order to save the information in the dataset, from that one is to confirm that the predictive ability of a feature is conserved. As a result, the organization results and related rules stay preserved, as well. To gather this requirement, the owner is supposed to define the usability constraints that provide the distortion band within which the principles of a feature can modify for each and every feature. As a result, the arrangement accuracy of the dataset remains unchanged. In addition to this, an added watermark should be unnoticeable and robust against any type of complicated attacks that can be launched on the watermarked dataset. And finally to defining usability constraints is a big challenge because a client has to maintain a balance between robustness of the watermark and preserving knowledge contained in features. For example, biomedical datasets may accept only very small amount of change during the embedding of a watermark in their features set to preserve the diagnosis rules. At the moment, the process of defining usability constraints is generally repeated and generally depends on the dataset and its intended application. Also if the right protection is forced using fingerprinting, the title-holder of data may need to define different usability constraints on the same dataset because in fingerprinting some different watermark for each user is added. From an information, no one method has been projected to model the usability


DOI 10.17148/IJARCCE.2015.47111

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 4, Issue 7, July 2015*

constraints for watermarking data mining datasets in particular, and also other relational datasets in general.[1]

## II. RELATED WORK

From our knowledge, no one technique has been planned for modelling usability constraints for watermarking data mining datasets. But there are some techniques define bellows

### *A. Technique of Watermarking Numeric Attributes in a Database*

In the work of Agrawal et al. the first well famous technique for watermarking numeric attributes in the database has been proposed. In that technique, message authenticated code (MAC) is calculated by helping of a secret key to identify the candidate tuples.[6]

1. In these, it is one of the Watermarking techniques of numerical data.
2. This Technique is highly dependent on a secret key.
3. It Uses markers to trace tuples to hide watermark bits. And it hides that watermark bits in the least significant bits.

Disadvantages

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

### *B. Rights Protection for Relational data*

Sion et al. presented marker tuples based watermarking technique for relational databases, but these techniques are inapplicable to data mining datasets because they do not aim at preserving the knowledge contained in the dataset. Protecting rights over relational data is of still increasing interest, especially considering areas where responsive, valuable content is to be outsourced. A simple and better example is a data mining application, where data is put up for sale in pieces to parties specialized in mining it. Different opportunities for right's protection are presented, each one with its own advantages and drawbacks. Enforcement by authorized means is usually not so effective in preventing theft of patented works, unless improved by a digital counter-part, for example watermarking.

The main reason of Digital Watermarking is to keep certain content from unauthorized duplication and sharing by enabling provable ownership over the content. It has conventionally relied upon the availability of a great noise domain within which the object can't be unchanged while retaining its essential properties A big challenge of watermarking is to embed an indelible mark in the object such that

1. The insertion of the mark does not destroy the cost of the object that is the object is still useful for the intended purpose and
2. It is difficult for an adversary to eliminate or alter the mark away from detection without destroying the value of that object.

Disadvantages

1. A model of attacks in this new-fangled domain needs to be devised.
2. More in depth attack analysis should be performed.
3. Adding novel data to the set is difficult.

### *C. Watermarking Technique For a Partitioning Based Database:*

Shehab et al. projected a partitioning based database watermarking technique. They modeled the process of watermark embedding as a constraint optimization problem and tested genetic algorithm (GA) and pattern search (PS) optimizers. They select PS because it is capable to optimize in concurrent But this method requires defining usability constraints manually and does not account for preserving the knowledge enclosed in the data mining datasets.[8,]

### *D. Information-Preserving Watermarking system for Right Protection of EMR System:*

Recently, in, we have proposed a relevant technique protecting ownership of electronic medical records (EMR) system. In this technique, information increase is used to identify the predictive ability of all features present in the EMR. The numeric features with the smallest amount predictive ability are selected to embed watermark bits to ensure information-preserving characteristic. This technique is only incomplete to information gain and does not generalize to other feature selection schemes. Furthermore, it does not take into account certain characteristics of dataset that play a very important role in the classification of the dataset. Since the major inspiration of the technique is information preserving watermarking; therefore, it does not explain any mechanism to model the usability constraints. Also, this watermarking technique is incomplete to numeric features only.[9]

Disadvantages:

In this technique, information gain is used to identify the predictive ability of all features there in the EMR. The numeric features with the least predictive ability are chosen to embed watermark bits to ensure information-preserving feature.

This technique is only limited to data gain and does not generalize to other feature selection schemes. Likewise, it does not take into account certain characteristics of dataset that play a fundamental role in the classification of the dataset. The main motivation of the technique is information-preserving watermarking. So, it does not describe any mechanism to model the usability constraints. in addition, this watermarking technique is imperfect to numeric features only.

.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 4, Issue 7, July 2015*

## III. WORKING MODEL

In this, there are two contributions:

1. a novel framework model which derives usability constraints for each kinds of datasets; and
2. a new watermarking technique that facilitates for numeric, non-numerical and strings datasets.

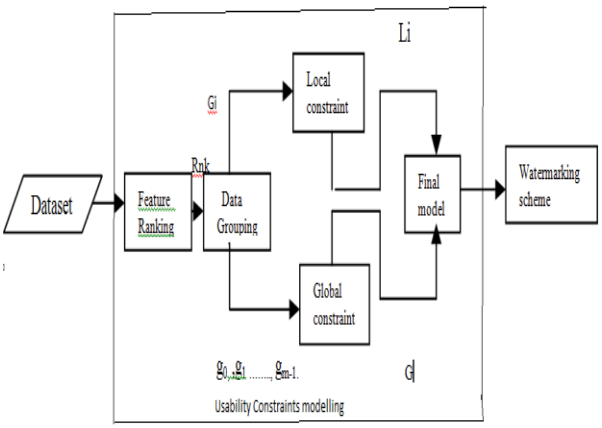


Fig.1 usability constraint model

This system takes the dataset as an input, models the usability constraints to be enforced through the watermark embedding in the dataset. Then it uses three different types of optimizers to find an optimum watermark that meets the relative constraints. Fig.1 shows the top level architecture of the proposed framework.

In the first step, the predictive ability of features, present in the dataset, are first calculated and then features are ranked on the basis of computed predictive ability Using these feature ranks, the next step is to form the logical groups of features. In this step, local usability constraints are defined for each and every logical group. Similarly, the global usability constraints are also defined that are applicable for the entire dataset. Finally, both types of constraints are used to construct a meta-constraints model that is given as an input to the watermarking scheme.

### *A. Watermarking Scheme:*

Watermarking scheme with its base in the above-mentioned formal model that not only preserves the classification potential of features but also results in approximately zero information loss. There are two main phases in these watermarking schemes: watermark encoding and watermark decoding.

- Watermark Encoding :

There are some steps involved in the watermark encoding phase that are

Step 1: The classification potential of each and every feature is calculated using mutual information I. and it is stored in the vector rank. The threshold $_{Cpt}$ is calculated using a vector of classification potentials. The classification potential of features, the vector and are then used to logical groups features of the dataset into the non-overlapping groups.

Step 2: The watermark is optimized and inserted in this stage while implementing the usability constraints model.

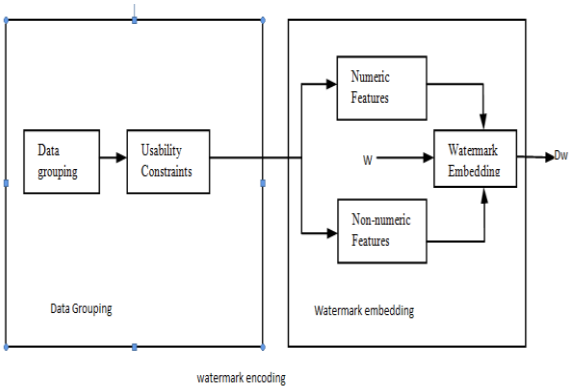


Fig.2 watermark encoding

- Watermark Decoding **:**

The architecture of the watermark decoding phase is shown in Fig.3.

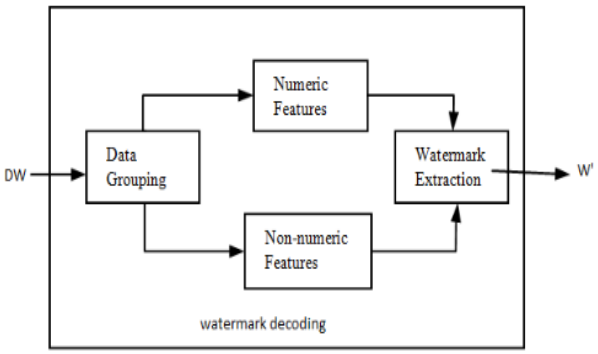


Fig 3. Watermark decoding

### *A. Watermark Decoding From Non-numerical Features:*

In the watermark detection phase, the hash value of a feature for each and every row is calculated by using the same steps of watermark embedding. The secret ordering, on the basis of this hash value, is calculated by comparing it with the ordering based on the hash values in temp and also the embedded bit is decoded from this ordering. It is important to state here that the sole purpose of using temp for ordering the rows during watermark decoding stage is to combat attacks which may vary the order of rows in the watermarked data. Since, the temp contains the ordering after embedding the last bit of that watermark; therefore, the last embedded bit is decoded first.

### *B. Watermark Decoding from Numeric Features:*

To decode the watermark from numeric features, a decoding threshold $T^*$ value is calculated for each group by using the same technique as in [10]. For decoding that watermark from a numeric feature in a group $G_{iw}$, a parameter is calculated from the watermarked dataset Dw as:

$$Val = \Delta i * G_{iw}$$

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 4, Issue 7, July 2015*

The value of the parameter Val. is compared with the decoding threshold and if Val. is established to be greater than $T^*$, then the watermark bit is decoded as 1; otherwise it is decoded as 0. Here, as an error correction mechanism, a majority voting step is performed after decoding a watermark bit from each row.

## IV. CONCLUSION

In this paper, a novel knowledge-preserving and lossless usability constraints model and a new watermarking scheme has been planned for watermarking data mining datasets. The payback of our techniques are: (1) recognizing the vital uniqueness of a dataset which require to be preserved at the time of watermarking; (2) ranking the features on the source of their classification potentials; (3) logically grouping the data into different groups or clusters based on this ranking for describing local usability constraints for each and every group; (4) defining global usability constraints for the whole dataset; (5) modeling the local and global usability constraints in such a manner so that the learning statistics of classifiers are conserved; (6) optimizing the watermark embedding such that all usability constraints remain unbroken; (7) make sure watermark protection by using data grouping and top secret parameters. As our knowledge, no technique in the literature exists that automatically compute "usability constraints" for a dataset that once enforced would preserve the knowledge contained in it. in addition, the enhanced watermarking scheme can work with any type of data: that is numeric and non-numerical with more watermark security.

## REFERENCES

[1] M. Pinsker, Information and Information Stability of Random Variables and Processes. San Francisco, CA, USA: Holden-Day, 1960
[2] Larry A. Dunning, Member IEEE, and Ray Kresman Privacy Preserving Data Sharing With Anonymous ID Assignment" IEEE Transaction on Information forensics and security , VOL. 8, NO. 2, February 2013, pp 402-413
[3] R. Agrawal, P. Haas, 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, M. Kwon, D. Ma, Q. Shao, and Y. Zhang, \Experience with software watermarking," in Proc. 16th Ann. 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 a proof-of-concept implementation, in Information Hiding. New York, NY, USA: Springer, 2001, pp. 185200.
[6] R. Agrawal and J. Kiernan, \Watermarking relational databases," in Proc. 28th Int. Conf. Very Large Data Bases, 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] M. Kamran andM. 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.
[10] M. Kamran and M. Farooq, A Formal Usability Constraints Model for Watermarking of Outsourced DataMining Datasets Tech. Rep. TR-59-Kamran, 2012

## BIOGRAPHY

**Miss. Rahinj Kanchan Shivaji** receiving her B.E. in Information Technology from Pune University, in 2012. She is currently pursuing her M.E. in Information Technology from University of Pune