Watermarking of Dataset with Usability Constraints Model

R. Shankari[*1], V. Sindhiya[*1], D. Vidhya[*1], Mrs.D.Shiny Irene[#2], Mrs.M.Arshiya Mobeen[#2]

[*1]*U.G Students, B.E CSE, Alpha College of Engg, Chennai, T.N, India.*

[#2]*Assistant Professor, Dept of CSE, Alpha College of Engg, Chennai, T.N, India.*

[1] advidhya@gmail.com

***Abstract* -** The large datasets are being mined to extract hidden knowledge and patterns that assist decision makers in making effective, efficient, and timely decisions in an ever increasing competitive world. This type of "knowledge-driven" data mining activity is not possible without sharing the "datasets" between their owners and data mining experts (or corporations); as a consequence, protecting ownership (by embedding a watermark) on the datasets is becoming relevant. The most important challenge in watermarking (to be mined) datasets is: how to preserve knowledge in features or attributes? Usually, an owner needs to manually define "Usability constraints" for each type of dataset to preserve the contained knowledge. The model aims at preserving "classification potential" of each feature and other major characteristics of datasets that play an important role during the mining process of data; as a result, learning statistics and decision-making rules also remain intact. We have implemented our model and integrated it with a new watermark embedding algorithm to prove that the inserted watermark not only preserves the knowledge contained in a dataset but also significantly enhances watermark security compared with existing techniques.

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

I. INTRODUCTION

The large datasets generated from very large database are being mined to extract hidden knowledge and patterns that are proving useful for decision makers to make effective, efficient and timely decisions in a competitive world.

This type of "knowledge-driven" data mining expert system cannot be designed and developed until the owner of data is willing to share the dataset with data mining experts. Recently, a startup company has made a business case out of this need where organizations outsource their datasets and the associated business challenge to data mining experts with an objective to find novel solutions to the posted problem .This validates the thesis that corporations with large databases want to get the optimized solution to a problem by levering the power of crowd-sourcing. In the emerging field of "sharing datasets" with the intended recipients, protecting ownership on the datasets is becoming a challenge in itself. Recently, an article reported the illegal sale of patients data and the concerned patients have sued the original hospital for breaching their privacy. An even bigger concern is that the recipient may try to take credit for contribution towards knowledge discovery and data mining by claiming the false ownership of the shared data. To mitigate these threats, a nondisclosure agreement is usually signed with the recipient binding him that he will not sale the dataset and will also not claim the ownership of the data. If the recipient breaches the agreement, the legitimate data owner can only sue him if he can prove in a court of Law his ownership over the dataset. Watermarking is the commonly used mechanism to enforce and prove ownership for the digital data in different formats like audio, video, image, relational database, text and software. The most important challenge in watermarking data mining datasets is: how to preserve knowledge in features or attributes during the embedding of watermark bits? In order to preserve the knowledge in the dataset, one has to ensure that the predictive ability of a feature or an attribute is preserved; as a result, the classification results remain preserved as well. To meet this requirement, an owner is supposed to define the "usability constraints" that provide the distortion band within which the values of a feature can change for each feature. As a result, the classification accuracy of the dataset remains unaltered. In addition to this, the inserted watermark should be imperceptible and robust against any type of sophisticated attacks that can be launched on the watermarked dataset. To conclude, defining "usability constraints" is a challenge because a user has to strike a balance between "robustness of watermark" and "preserving knowledge contained in features". For example, biomedical datasets may

tolerate 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 manually repeated and is dependent on the dataset and its intended application. Moreover, if right protection is enforced using "fingerprinting", the owner of data may need to define different "usability constraints" on the same dataset because in fingerprinting a different watermark for each user is added. To the best of our knowledge, no technique has been proposed to model the "usability constraints" for watermarking data mining datasets in particular, and other relational datasets in general. In this paper, we propose a novel formal model for identifying the essential "usability constraints" which must be enforced while embedding watermark in a dataset. A generic formal model to define "usability constraints" on a dataset that not only ensures the robustness of an inserted watermark is proposed but also preserves the knowledge contained in the dataset. The proposed technique is independent of the type of a dataset i.e. numeric or nonnumeric. It is a new knowledge-preserving watermarking scheme to validate its efficacy and effectiveness. The new knowledge-preserving watermarking scheme has significantly enhanced the security deleting or changing the watermark compared with existing techniques.

II. RELATED WORKS

To the best of our knowledge, no technique has been proposed for modeling "usability constraints" for watermarking data mining datasets. In the work of Agrawal[3], the first well known technique for watermarking numeric attributes in a database has been proposed. In this technique, message authenticated code (MAC) is calculated with the help of a secret key to identify the candidate tuples. Sion[7] presented a marker tuples based watermarking technique for relational databases but these techniques are not applicable to data mining datasets because they do not aim at preserving the knowledge contained in the dataset. Shehab[8] proposed a partitioning based database watermarking technique. They modeled the process of watermark insertion as a constraint optimization problem and tested genetic algorithm (GA) and pattern search (PS) optimizers. They select PS because it is able to optimize in real-time. But this technique requires defining "usability constraints" manually and does not account for preserving the knowledge contained in the data mining datasets. Recently, we have proposed a relevant technique protecting ownership of electronic medical records (EMR) system[10]. In this technique, information gain is used to identify the predictive ability of all features present in the EMR[10]. The numeric feature(s) with the least predictive ability are selected to embed watermark bits to ensure information-preserving characteristic. This technique is only limited to information gain and does not generalize to other feature selection schemes. Moreover, it does not take into account certain characteristics of dataset that play a vital role in classification of the dataset. Since the major motivation of the technique is information-preserving watermarking; therefore, it does not describe any mechanism to model the "usability constraints". Moreover, this watermarking technique is limited to numeric features only. In comparison, the focus of our current work is on developing a formal model to define "usability constraints" for watermarking of data mining datasets in such a way that the watermark is not only robust but the knowledge contained in the dataset is also preserved. Furthermore, we also provide a mechanism to logically group the dataset into groups such that high ranked features might also be watermarked during watermarking. This is a significant enhancement because if only low ranked features are watermarked, an attacker can launch malicious attacks on low ranked features only without compromising the data quality to a great extent. In this context, our data grouping approach enables a data owner to embed a watermark in high ranked features as well while still satisfying the "usability constraints" imposed by our formal model. Last but not the least, we have significantly enhanced our recently proposed information-preserving watermarking scheme for data mining datasets in such a way that it can now watermark any type of features numeric, nonnumeric.

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

• 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, nonnumeric, strings 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 feature.

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, two different constraints are used to watermark the dataset. They are Local usability constraints, Global usability constraints.

LOCAL USABILITY CONSTRAINTS

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

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.

ARCHITECTURE DIAGRAM

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

The classification potential features are used to logically group features of the dataset into no overlapping groups .The watermark is optimized and embedded to ensure the usability constraint modeled. clients simultaneously, and offer shared 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) Modeling 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.

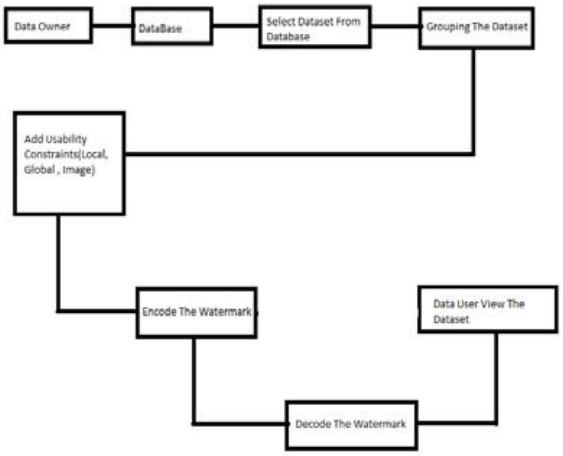

fig I.1. System Architecture

V. WATERMARKING SCHEME

A. WATERMARK ENCODING

Different steps are involved in watermark encoding phase [fig I.1]. 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.

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.

• Global constraints applied for the whole dataset.

• Local constraints applied for the logical group of the dataset.

• 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.

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.

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

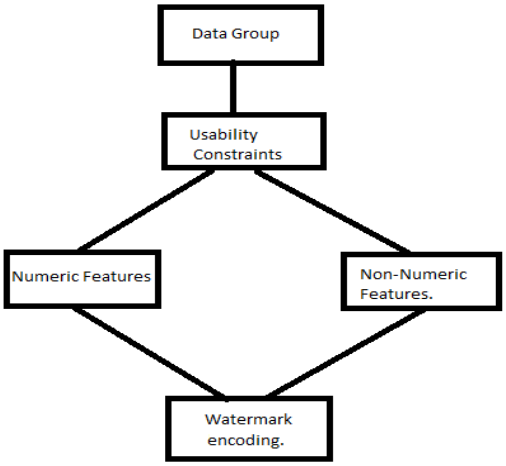

Fig I.2 *Watermark Encoding*

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 I.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.

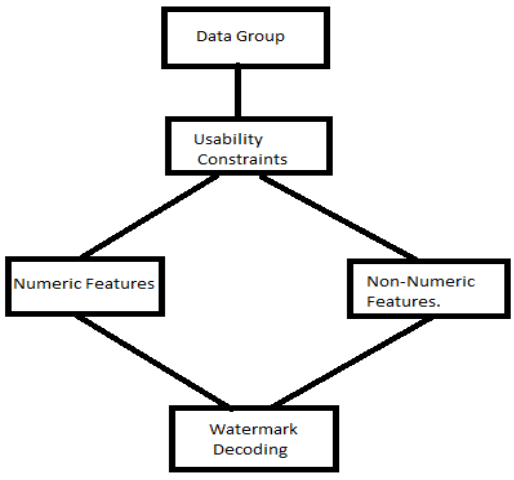

Fig I.3 Watermark Decoding

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.

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.

VII. REFERENCES

[1] Kaggle's Contests: Crunching Numbers for Fame and Glory 2012 [Online]. Available: http://www.businessweek.com/magazine/kaggls-contests-crunching-numbers-for-fame-andglory-01042012.html

[2] Patients Sue Walgreens for Making Money on Their Data 2012 [Online]. Available: http://www.healthcareitnews.com/news/patients suewalgreens-making-money-their-data

[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 InformationHiding. 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 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] 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

