Security–Preserving Data Mining: Watermarking Of Outsourced Datasets Using Usability Constraints Model

Lakshmi Priya U.S[1] and Soya Chandra[2]

[1]PG scholar, Dept. Of CSE, Sarabhai Institute Of Science & Technology, Trivandrum, India

[2] Assistant Professor, Dept. Of CSE, Sarabhai Institute Of Science & Technology, Trivandrum, India

***Abstract*— The knowledge driven data mining systems cannot be developed and designed until the owner of the data is willing to outsource the data with corporations or data mining experts or corporations. In the emerging field of outsourced datasets with the intended recipients, protecting ownership of the data is becoming a challenge in itself. The commonly used mechanism to enforce and proves ownership for the digital data in different formats is watermarking. How to preserve knowledge in features or attributes during the embedding of watermark bits is the most important challenge in watermarking relational databases. An owner usually needs to define usability constraints manually for each type of dataset to preserve the contained knowledge. Major contribution of this paper is a formal model that defines usability constraints for each dataset in an automated fashion. The classification potential of each feature and other major characteristics of dataset that play an important role during the mining process of data are preserved; as a result, decision making rules and learning statistics also remain intact. Also we introduce a new class of aggregate functions that aggregate numeric expressions and transpose results to produce a data set with a horizontal layout. In addition to this, the model is integrated with a new watermark embedding algorithm to prove that the inserted watermark should be imperceptible and robust against any type of attacks.**

***Index Terms*—Ownership preserving data mining, right protection, usability constraints, watermarking datasets.**

I. INTRODUCTION

In a knowledge driven data mining system, the datasets generate from large databases are mined to extract patterns and hidden knowledge that are useful for decision makers to make efficient, effective and timely decisions. This type of knowledge driven data mining systems cannot be designed and developed until the owner of the data is willing to outsource the data with data mining experts or corporations.

Recently a startup firm called "kaggle" outsource their datasets and the associated business challenge to data mining experts to find novel solutions for their posted problem [1]. This validates that the organizations with large databases needs to get optimized solutions to their problems by levering the power of crowd-sourcing. In the emerging field of outsourced datasets with the intended recipients, protecting ownership of the data is becoming a challenge in itself.

Watermark can be applied to any database relation that has 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:

- A model that automatically identifies the usability constraints on a dataset that ensures the robustness of inserted watermark.
- The proposed technique is independent for numeric and non numeric features of a dataset.
- A new knowledge preserving watermarking scheme is integrated with this model to validate its effectiveness and efficacy.

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

II. RELATED WORK

The work of R Agrawal, J Kiernan [3], is 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 is used to identify the candidate tuples. J Kiernan, P Haas, R Agrawal [3] proposed watermarking tuples in a relational database that uses signals. It inserts watermark with multiple bits on multiple tuples. But it needs to improve the optimization.

M Crogan, V Raskin, M Atallah [6], presented two main results in the area of information hiding in natural language text. This 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. Ghafoor, E Bertino, M Shehab [9], describes a partioning based watermarking technique. They consider 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.

S.krishnaswami [5], experimented with a watermarking system for java which embeds a watermark in dynamic data structures. It shows that watermarking can be done efficiently to gradually increase the size of code, heap space and execution times. R Sion, S Prabhakar [8], 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.

M Farooq, M Kamran [12], recently proposed a technique for protecting the ownership of EMR (Electronic Medical Records) system. In this work 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.

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.

The proposed work is focused 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 hacked by an attacker by launching malicious attacks. And we proposed watermarking for numeric and non-numeric attributes.

III. APPROACH OVERVIEW

The proposed work presents two contributions: i) a model which derives usability constraints for all kinds of datasets. ii) A new watermarking technique works for numeric and nonnumeric datasets. The 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.

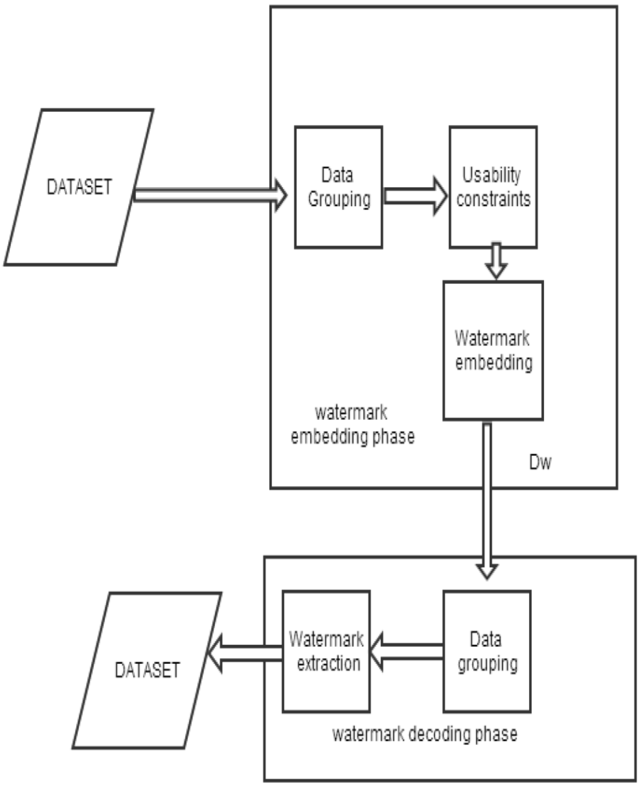

Fig.1.Architecture Diagram Of The Proposed System

A MODEL FOR DEFINING USABILITY CONSTRAINTS

This proposed model defines the Usability Constraints, which is used to preserve the data during the process of watermark insertion in the dataset. This constraints provides a distortion band within which the values of a feature can change for each feature.

The model takes the dataset as input and defines the “usability constraints’ to be enforced during the watermark embedding in the dataset. First step is to calculate the predictive ability of each feature present in the dataset and based on this the features are ranked. These ranks are used to generate the logical groups of features in the next step. Here, local usability constraints are defined for each logical group and the global usability constraints are also defined that are applicable for the whole dataset. Finally, both types of constraints are combined to build a meta-constraints model. This combined constraints model is given as an input to the watermarking scheme.

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

Definition: 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.

Definition: watermark embedding

Watermark embedding is a transformation of a dataset Do to Dw after embedding the watermark W.

Definition: Feature selection scheme

A feature selection scheme S transforms M-dimensional data DO, having N samples, M features and a class attribute Y , in m-dimensional space Rm (with m ≤ M, such that Rm ⊆ RM) that can yield“optimum” learning statistics.

In this paper, we have used 6 most commonly used different feature selection schemes (mutual information (I), information gain (IG), information gain ratio (IGr), correlation based feature selection (CFS), consistency based feature subset evaluator (CBF), and principal components analysis (PCA)). All these feature selection schemes define the classification potential CP of the features.

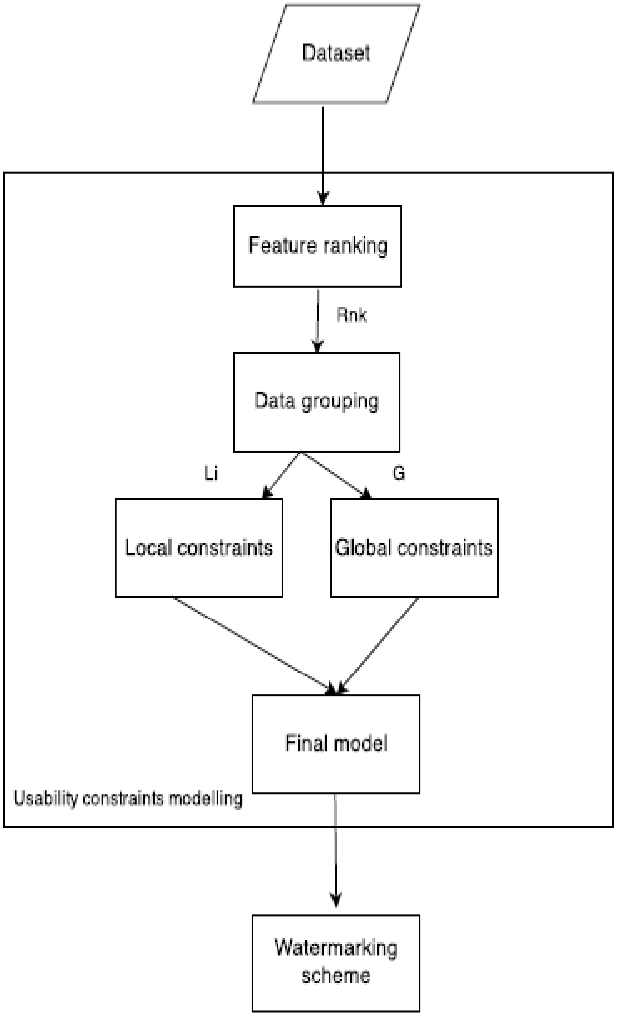

Fig.2. Top level architecture of the proposed system.

IV. WATERMARKING SCHEME

The foundation of the watermarking scheme is the proposed model for usability constraints. There are two main phases in this watermarking scheme: watermark encoding and watermark decoding.

A. *Watermark Encoding*

The steps involved in this phase are:

1) *Feature Ranking:* Logically group the data into' n' nonoverlapping partitions and define the usability constraints whose information loss is zero. Features are ranked using mutual information and these ranks are stored in a vector Rnk.

2) *Classification Potential Computation:* It is important to compute the amount of change that a feature can tolerate during the watermarking process. The features with high classification potential can tolerate only small changes and the top ranked features shows zero tolerance towards any change

3) *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. The groups are used 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

4) *Refined Usability Constraints:* Refine the usability constraints into two 1)Global constraints applied for the whole dataset,2)Local constraints applied for the local group of the dataset.

5) *Select Data for Watermarking*: The selection of relevant rows for embedding watermark is the important step in watermarking of a dataset. A parameter is used to store the information about the selected rows. Its main purpose is to insert a watermark for the selected rows to preserve the data presented in the dataset

6) *Watermark Embedding:* It involves two phas*es*

a) 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. A sequence of binary bit is used to embed watermark in a dataset. Secret hash value for each row is computed by seeding a pseudo random sequence generator and it is concatenated with a class label of the row, secret key and row value. The secret order does not make any change in the underlying dataset. Same hash value would be generated if a row value is repeated with same class label. After the embedding of final bit (Dw), the secret order of hash value is stored to use it during the decoding Dw

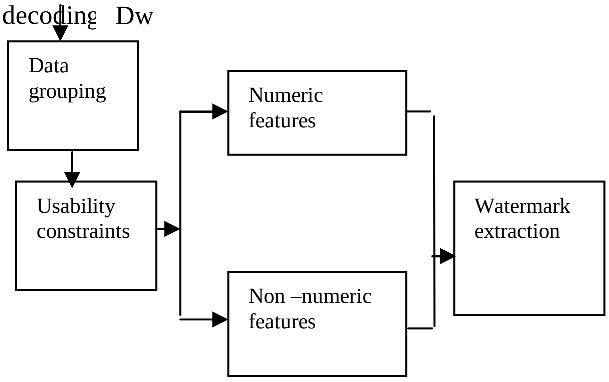

Fig.3. Watermark Decoding Phase

b) 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 alteration and with bit 0 as negative alteration.

B. WATERMARK DECODING

1) Watermark Decoding From Non-Numeric Features

The watermark decoding is the reverse process of watermark encoding. For each row the hash value of a feature is computed using the same steps of watermark embedding. This hash value is used to calculate the secret ordering and based on that embedded bit is decoded.

2) Watermark Decoding From Numeric Features

Watermark decoding from numeric features computes a decoding threshold value using 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.

V. CONCLUSION

In this paper, we proposed a new method to define a usability constraint to preserve the knowledge contained in the dataset and it is integrated with a new watermarking scheme. The benefits of our techniques are:

- Features are ranked based on their computed classification potential.
- The modeling of constraints maximizes the lossless data.
- Preserve the knowledge contained in the dataset.
- It is difficult for the intruders to extract watermark.
- A new approach of modeling “usability constraint” is defined to preserve the dataset.
- Watermark security is ensured by the use of secret parameters and data grouping.
- Optimize Watermark embedding to ensure that the usability constraints remain intact.
- Enhanced the watermark technique from numeric features to non-numeric features with more watermark security

To the best of my knowledge, 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.

REFERENCE

[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://www.healthcarenews.com.

[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] ZHU Qin, YANG Ying, LE Jia-jin, LUO Yishu,(2006)"Watermark based Copyright Protection of Outsourced Database," IEEE, IDEAS, pp. 1-5.

[5] 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

[6] 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. Springer, 2001, pp. 185–200.

[7] R. Agrawal and J. Kiernan, "Watermarking relational databases," in 28th International Conference on Very Large Data Bases. Morgan Kaufmann Pub, 2002, pp. 155–166.

[8] R. Sion, M. Atallah, and S. Prabhakar, "Rights protection for relational data," IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 6, pp. 1509–1525, 2004.

[9] M. Shehab, E. Bertino, and A. Ghafoor, "Watermarking relational databases using optimization-based techniques," IEEE Transactions on Knowledge and Data Engineering, vol. 20, no. 1, pp. 116–129, 2008.

[10] 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.

[11] M. Kamran and M. Farooq, "A formal usability constraints model for watermarking of outsourced data mining datasets," IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, VOL. 8, NO. 6, JUNE 2013.

[12] 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