A Recognized Usability Limitations Model for Watermarking of Outsourced Datasets

Mr. G. Bhuvanendran [1] Mr. G. Karpagakannan[2]
[1]M. E. (Final Year) [2]M. E. & Assistant Professor
[1, 2]RatnaVel Subramaniam College of Engineering & Technology

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 discovery driven" data mining activity is not possible without sharing the "datasets" between their owners and data mining experts as a consequence, protecting. Usually, an owner needs to manually define "Usability limitations" for each type of dataset to preserve the contained knowledge. The major contribution of this paper is a novel formal model that facilitates a data owner to define usability limitations to preserve the knowledge contained in the dataset in an automated fashion. 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. We have tested our model on more than twenty five different data mining datasets to show its efficacy, effectiveness, and the ability to adapt and simplify.

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

I. INTRODUCTION

The large datasets generated from very large databases 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 based driven" data mining expert systems cannot be designed and developed until the owner of data is willing to share the dataset with data mining experts. Recently, a startup company "gaggle" 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 the datasets" with the intended recipients, protecting ownership on the datasets is becoming a challenge in itself. Recently, an article reported the illegal sale of patient's 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 (KDD) 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 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 limitations" 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 limitations" 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.

The major contributions of our paper are to be following:

- We propose a generic formal model to define "usability limitations" on a dataset that not only ensures the robustness of an inserted watermark but also preserves the knowledge contained in the dataset. The proposed technique is independent of the type of a dataset.
- We have integrated our model in a new knowledge-preserving watermarking system to validate its efficacy and effectiveness.
- We show that the new knowledge-preserving watermarking system has significantly enhanced the security deleting or changing the watermark compared with existing techniques.
- We have conducted experiments more than twenty five publicly available datasets to prove that our technique can generalize to any type of dataset and achieve its objective of preserving the classification accuracy when mined with a machine learning classifier.

II. RELATED WORK

To the best of our knowledge, no technique has been proposed for modeling "usability limitations" for watermarking data mining datasets. The first well known technique for watermarking numeric attributes in a database

has been proposed. In this technique, MAC = Message Authenticated Code is calculated with the help of a secret key to identify the candidate tuples. Sion et al. 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 et al. proposed a partitioning based database watermarking technique. They modeled the process of watermark insertion as a constraint optimization problem and tested genetic algorithm (shortly called GA) and pattern search (shortly called PS) optimizers. They select PS because it is able to optimize in real time. But this technique requires defining "usability limitations" manually and does not account for preserving the knowledge contained in the data mining datasets.

We have proposed a relevant technique protecting ownership of Electronic Medical Records (shortly called EMR) system. In this technique, information gain is used to identify the predictive ability of all features present in the EMR. The numeric feature with the least predictive ability is 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 systems. 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 limitations".

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 limitations" 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 limitations" imposed by our formal model. Last but not the least, we have significantly enhanced our recently proposed information based preserving watermarking system for data mining datasets in such a way that it can now watermark any type of features numeric or nonnumeric.

III. Proposed Approach

We present two contributions: 1. a novel framework model which derives usability limitations for all kinds of datasets. 2. A new watermarking technique that works for numeric, nonnumeric and strings datasets. Our system takes the dataset as an input, models the "usability limitations" 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 limitations. Fig. 1 shows the top level architecture of the proposed framework. In the first step, 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 limitations" are defined for each logical group. Similarly, the "global usability limitations" are also defined that are applicable for the whole dataset. Finally, both types of limitations are used to build a Meta limitations model that is given as an input to the watermarking system.

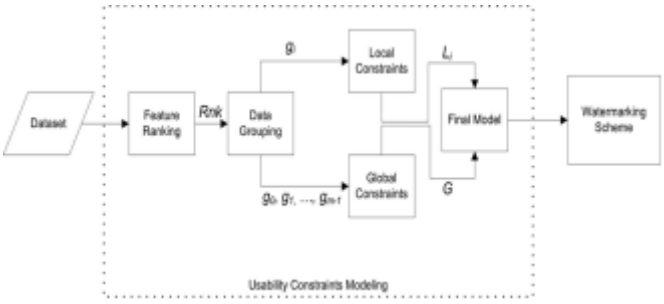

Fig. 1: Top level architecture of the proposed framework.

IV. A Prescribed Model for Usability Limitations

We now present our formal model to define "Usability Limitations" that preserve the knowledge during the process of inserting watermark in the dataset.

Definition1: Tuple: A tuple is an ordered list of elements. The tuple is used as a basic unit for referring different parameters of a dataset.

Definition 2: Learning Algorithm: Given a dataset with features, instances, and a class attribute, a learning algorithm, group's instances into different groups.

Definition3: Learning Statistics: Learning statistics is a tuple containing the classification statistics of a particular learning algorithm.

Decision rule boundaries: denotes the threshold values that define the boundary of a particular decision rule.

True Positive (TP): TP denotes the number of instances of a particular class detected as instances of that class.

False Positive (FP): For a particular class, the number of instances of other classes detected as instances of that particular class.

True Negative (TN): For a particular class, the number of instances detected as instances of other classes.

False Negative (FN): For a particular class, the number of instances of that class detected as instances of other classes

Definition 4:

Decision Rules: Given a dataset with 'Do', 'M' features, a rule 'R' is a tuple constructed by mapping of M features.

V. Watermark System

We now describe our watermarking system with its foundation in the above-mentioned formal model that not only preserves the classification potential of features but also results in zero information loss. There are two main phases in our watermarking system: watermark encoding and watermark decoding.

A. Watermarking Encoding:

The steps involved in the watermark encoding phase are: 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 non overlapping groups. Step 2: The watermark is optimized and embedded in this stage while enforcing the usability limitations modeled in Section IV. The different steps of watermark encoding phase are shown in Fig. 2.

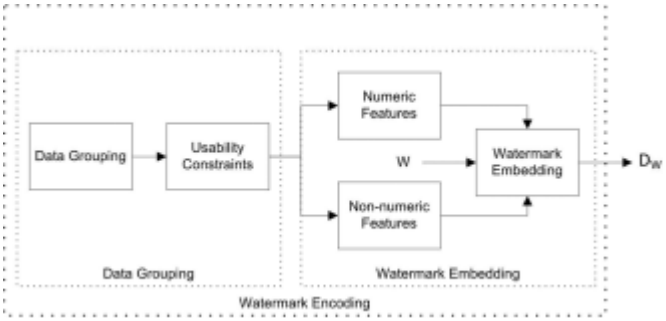

Fig. 2: Different steps of watermark encoding phase.

1) Feature Ranking:

In this step, the features are ranked to: No.1 logical groups the data into non overlapping partitions; and No.2 to define "usability limitations" in such a manner that the information loss is zero. The ranking is done, using a well-known information measure, mutual information, to understand the correlation of a feature on predicting a class label. The rank of all features, present in a dataset, is stored in a vector.

2) Classification Potential Threshold Computation:

Intuitively speaking, a feature that has a large classification potential is expected to tolerate only a small change in its values to ensure that the decision rules remain unaltered.

The features with high classification potential can tolerate Fig. 2. Different steps of watermark encoding phase only small changes, if the information is to be preserved during watermarking; and the top ranked features show approximately zero tolerance towards any change. Therefore, it is very important to compute the amount of change that a feature can tolerate during the watermarking process. The data groups are constructed using and "tolerable alteration" is computed for each group.

3) Data Grouping:

As mentioned before, is used to group the features into logical non overlapping groups.

The grouping function is applied on every feature of a given dataset. Group might be empty but then will be omitted during the optimization phase. Remember that the groups are only logical and hence cannot be truly separated from one another. Our pilot studies show that the proposed group assignment algorithm is simple yet effective.

We use the groups to define the local and global "usability limitations". Our data grouping approach overcomes a significant shortcoming of our earlier work in which only the lowest ranked features are watermarked. In the new approach, an attacker cannot easily build an attack vector by filtering low ranking features only.

4) Refined Usability Constraints:

In this paper, the data usability limitations are defined by "knowledge based preserving and lossless usability limitations model". In order to easily enforce limitations, we refine the constraint number into two types:

(1) global limitations for the whole dataset; and

(2) local for a particular logical group.

B. Local Usability Limitations

The local usability limitations are defined by mutual information. In order to enforce them, the constraint 2 of "knowledge based preserving and lossless usability limitations model" must be met. In our formal model, we preserve the information by trying to keep the data distribution unaltered.

As a result, we have to model the challenge as an optimization problem to ensure that under given limitations, the "tolerable alteration" is maximized for all features in general and minimized for high ranking features within each group while watermarking a group.

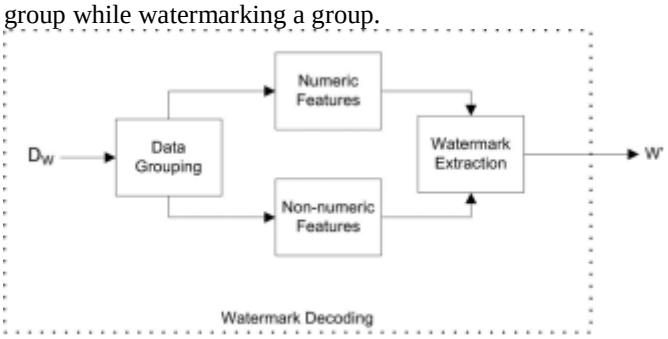

Fig. 3: Steps of watermark decoding phase.

VI. Experimental Result

We have performed our experiments on twenty five different datasets. These biomedical and biomedicine datasets are carefully chosen from different domains so that we test our technique for two class datasets, multiclass datasets, high dimensional datasets, datasets with missing values, datasets with various type of features, imbalanced datasets, and large datasets.

We also report the results of our study on the robustness of proposed scheme as compared to show that our approach in this paper improves upon a very important aspect of related to watermark security. The experiments were carried out on a computer with a one lakh seventy three Core II processor and a One Giga Bytes of RAM. We set the value of C=100 Percentage, which means that all rows were selected for watermarking. A data owner can choose any watermark length L = 16 bits but we set its length. Five well known machine learning schemes are used to analyze their learning statistics on both original and altered datasets to show the relevance of Theorem two in the proposed scheme.

VII. Conclusion

In this paper, a novel knowledge based preserving and lossless usability limitations model and a new watermarking scheme has been proposed for watermarking data mining datasets. The benefits of our techniques are: No.1 Identifying the vital characteristics of a dataset which need to be preserved during watermarking; No.2 Ranking the features on the basis of their classification potentials; No.3 Logically grouping the data into different groups based on this ranking for defining local usability limitations for each group; No.4 defining global usability limitations for the complete dataset; No.5 Modeling the local and global usability limitations in such a manner so that the learning statistics of a classifiers are preserved; No.6 Optimizing the

watermark embedding such that all usability limitations remain intact; No.7 Ensuring watermark security by using data grouping and secret parameters. To the best of our knowledge, no technique in the literature exists that automatically computes "usability limitations" for a dataset that once enforced would preserve the knowledge contained in it. Moreover, the enhanced watermarking scheme can work with any type of data: numeric and nonnumeric with more watermark security. The proposed technique can be easily employed by the customers of companies like Kaggle to share datasets with data-mining experts by safeguarding and protecting their ownership. The technique, in a future work, could be extended to images.

REFERENCES

[1] Patients Sue Walgreens for Making Money on TheirData2012.Available:http://www.healthcareitn-news.com/news/patients-suewalgreens-making-money-their-data
[2] Kaggle's Contests: Crunching Numbers for Fame&Glory2012.Available:http://www.businessweek.com/magazine/kaggles-contests-crunching-numbers-for-fame-and-glory-1042012.html
[3] R.Agrawal,P.Haas,and&J.Kiernan,"Watermarking relational data: Framework, algorithms and analysis," 2003.
[4] J. Palsberg, S. Krishnaswamy, M. Kwon, D. Ma, Q. Shao, and Y. Zhang, "Experience with software watermarking,"
[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,".
[6] R. Agrawal and J. Kiernan, "Watermarking relational databases,".

