A Formal Model to Preserve Knowledge in Outsourced Datasets

[1]Veera Ragavan K, [2]Karthick S
[1]ME Student, [2] Asst.Prof
[1,2]Dept of Software Engineering, SRM University, Chennai, India
[1]ragu.skp@gmail.com , [2], Karthik.sa@ktr.srmuniv.ac.in

***Abstract* - Large datasets are being outsourced with help of data owners and mining experts. This type of 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. In general owner need to define the usability constraints manually to preserve the knowledge contained in the dataset. This paper aims at creating a framework to define the usability constraints in an automated fashion. This novel formal model facilitates a data owner to define usability constraints and to preserve the knowledge contained in the datasets in automated fashion and to provide security to outsourced dataset. We implemented and tested our model on different data sets .Our model not only preserve knowledge contained in the datasets but significantly enhances security to datasets compared with existing system**

***Keywords* - Data usability, Security, Knowledge-preserving, data mining, Usability constraints.**

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-driven" data mining expert cannot be designed and developed until the owner of data is willing to share (or outsource) the dataset with data mining experts. Recently, a startup company Kaggle 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 [1]. 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[2].In order to preserve the knowledge in the dataset, one has t 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 At the moment, the process of defining "usability constraints" is manually repeated and is dependent on the dataset and its intended application.

To the best of our knowledge, one technique[9] has been proposed to model the "usability constraints" for outsourcing datasets . In this paper, we propose a novel formal model for identifying the essential "usability constraints" of the datasets and security to the outsourced dataset.

The major contributions of our paper are:

- We propose a generic formal model to define "usability constraints" on a datasets in an automated fashion . The proposed technique is independent of the type of a dataset i.e. numeric or nonnumeric.
- We implemented basic encryption decryption mechanism to datasets

We briefly describe the related work in Section II and then introduce our approach in Section III. We present the formal model in Section IV. Subsequently, we explain that how the "usability constraints", defined by our model, We report the results of our experiments in Section VI and conclude in final section

II. RELATED WORK

Early one technique has been proposed for modeling usability constraints in an automated fashion[9] for datasets. In the work of Agrawal et al. [3], the first well known technique 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 et al. [4] presented a marker tuples based 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. [5] proposed a partitioning database technique. They modeled the process as a constraint optimization problem and tested genetic algorithm (GA) and pattern search (PS) [6] 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 containedin the data mining datasets.

Recently in [7],they have proposed a relevant technique protecting ownership of electronic medical record system this information gain is used to identify the predictive ability of all features present in the EMR .The numeric feature with least predictive ability are selected to outsource datasets and this technique is limited to information gain and does not generalize to other feature selection schemes..Moreover it does not take onto account certain characteristics of datasets that play a vital role in classification of datasets.

In comparison of existing system [8][9] our focus is on developing a formal model to define usability constraints in an automated fashion. We also provide a mechanism to logically group the datasets into group such that low ranked and high ranked features are outsourced so an attacker cannot launch attack on datasets. This technique allows data owner to outsource low and high ranked features and also define usability constraints in automated fashion

III. APPROACH OVERVIEW

In this paper, we present two major contributions: (1) a formal model to define usability constraints for all kinds of datasets in an automated fashion; (2) a security protection to outsourced datasets using simple encryption decryption algorithm. Our system takes input as datasets and models the usability constraints for datasets to outsourced .Later it uses simple encryption mechanism to provide security to datasets.

In the first step the predictive ability of features present in the datasets are calculated and the features are ranked on the basis of computed predictive ability .these ranks are used to generate the logical groups of features .In this step the local usability

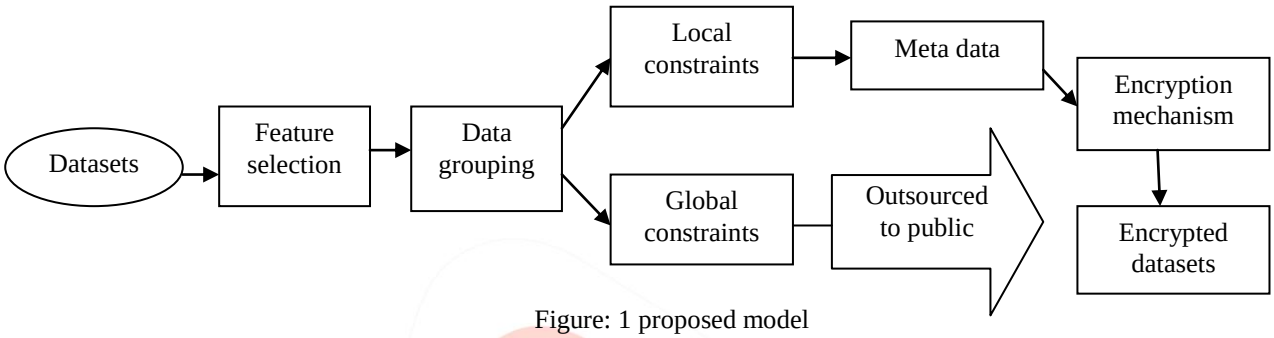

Figure: 1 proposed model

Constraints area defined for each logical group .Similarly, the global usability constraints are also defined that are applicable for whole datasets. Finally both types of constraints are used to build Meta data constraints model that is given as input to the encryption. Simple encryption mechanism is used which takes the final model as input and encrypt it which is outsourced to the mining expert later.

IV. USABILITY CONSTRAINTS MODEL

We now present our formal model to define usability constraints that preserve the knowledge in datasets which is being outsourced. Tuple is τ an ordered list of elements this tuple is used as a basic unit for referring different parameters of a dataset Learning Algorithm may be a classification algorithm or a clustering algorithm .Given a datasets with M features ,N instances and class attribute Y, a learning algorithm Γ, groups N instance into α groups.Learning statistics C_s is a tuple containing the classification statistics of a particular learning algorithm these statistics include true positive, false positive and decision rules.

In this paper, we have used 6 most commonly used different feature selection schemes mutual information, information gain ,information gain ratio , correlation based feature selection, consistency based feature subset evaluator, and principal these feature selection schemes define the classification potential of the features Classification Potential Threshold is used to make different groups of a dataset so that the features with higher classification potentials are least modified during outsourced Local usability constraints L_i is a tuple constituting mutual information of a feature in a particular group .These local usability constraints are used to outsource a features in a group and they are enforced at a group level only.Global usability given a dataset global usability constraints G is a tuple that consists of features set produced by different feature selection schemes on a that datasets in our case we are using 5 different feature selection schemes . Global usability constraints are enforced both at a group level and at the global dataset level. The features' set, produced by applying a feature selection scheme to a group or a dataset, should remain unaltered. The local usability and global usability constraints , learning statistics .

V. ENCRYPTION/DECRYPTION MECHANISM

We are using simple cryptography algorithm for encryption decryption mechanism of outsourced datasets

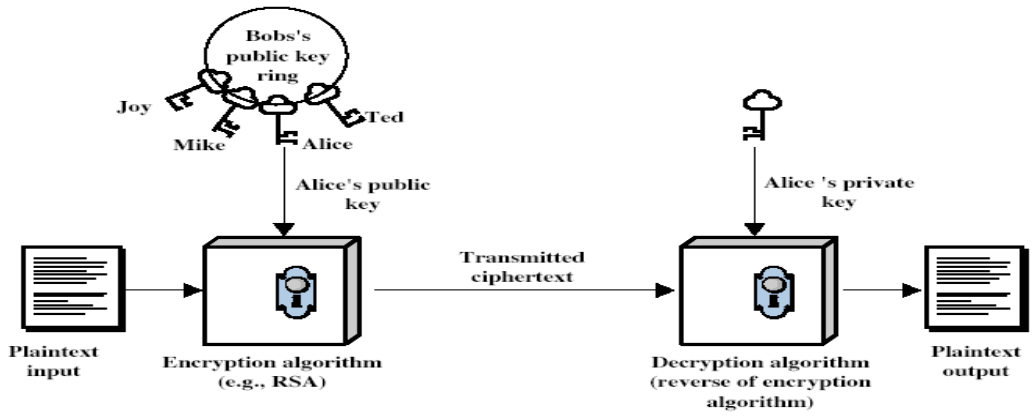

Figure:2 RSA Architecture diagram

RSA algorithm uses both public and private key for encryption and decryption of datasets it looks simple but has two layer

protection since it uses two keys for encrypt and decrypt The mathematical details of the algorithm used in obtaining the public and private keys are available at the RSA Web site. Briefly, the algorithm involves multiplying two large prime numbers and through additional operations deriving a set of two numbers that constitutes the public key and another set that is the private key. Once the keys have been developed, the original prime numbers are no longer important and can be discarded. Both the public and the private keys are needed for encryption /decryption but only the owner of a private key ever needs to know it. Using the RSA system, the private key never needs to be sent across the Internet.

The private key is used to decrypt text that has been encrypted with the public key. Thus, if one send a message, The receiver can find out your public key , from a central administrator and encrypt a message to receiver using your public key. When the receiver receive it, receiver decrypt it with sender private key. In addition to encrypting meassages, the receiver can authenticate himself to sender by using receiver private key to encrypt a digital certificate. When the sender receive it the sender can use receiver public key to decrypt it.

[10]RSA algorithm can be classified as three algorithms, the key generation algorithm, encryption algorithm, and decryption algorithm.

RSA key generation algorithm can be described as follows

1. Generate two large random and distinct primes P and Q
2. Calculate N = P.Q and □□= (P – 1)(Q – 1)
3. Choose a random integer E, 1 < E < □, such that gcd(E,□) = 1
4. Compute the unique integer D, 1 < D < □, such that ED ≡□ 1 (mod □)
5. Public key is (N, E) and private key is (N, D) RSA encryption algorithm can be described as follows,

C = ME mod N, RSA decryption algorithm can be described as follows,

M = CD mod N, which C represents ciphertext and M represents message.

VI. EXPERIMENT AND RESULTS

We have performed our experiments on 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. Five well known machine learning schemes are used to analyze their learning statistics on both original and altered datasets to show the relevance in the proposed scheme. We also report the results of our study on the robustness of proposed scheme as compared to [7][8][9] to show that our approach in this paper improves upon a important aspect related to securityThe experiments were carried out on a computer with a 1.73 Core 2 processor and a 1 GB of RAM.

A. Knowledge Preserving Characteristic Of Proposed Model

We give the optimized "usability constraints", derived from our formal model, to the enhanced encryption algorithm. We report the effect of encryption, with the proposed usability constraints model, on various feature selection schemes in Table I (for brevity we report our result on only one dataset) it is proven that mutual information between a feature and the class attribute can only remain same for original and encrypted datasets, if the features belong to the same class in both datasets. we report mutual information of original and encrypted features in Table I to prove that the features' values of the encrypted dataset have the same relation with the class labels as they had before encryption It is evident from Table I that learning statistics of all feature selection schemes have also been preserved by enforcing "usability constraints" of our formal model. Once the dataset is encrypted, we classify—using five well known machine learning algorithms—all 25 original datasets and their corresponding 24 encrypted datasets. The learning statistics are preserved for all learning algorithms with the exception.But we can safely generalize that the overall learning statistic preserved by enforcing the usability constraints of our formal model. Table I shows the learning statistics of original and encrypted dataset For brevity, we show in [8] the classification rules, extracted by J48 only, for the original and encrypted datasets. It is evident that the rule boundaries the third element of learning statistic are also preserved and the rules to predict the class labels have remained unchanged.

Table I Effect Of Encryption On Various Feature Selection Schemes. S Denotes Feature Selection Scheme

ORGINAL DATA					
S	F1	F2	F3	F4	F5
I	.062	.304	.059	.082	.277
I_G	.039	.190	.014	.044	.06
IG_r	.052	.099	.014	.022	.04
CFS	No	Yes	No	No	No
CBF	Yes	Yes	Yes	Yes	Yes
PC	.738	.522	.393	.284	.188
ENCRYPTED DATA					
S	F1	F2	F3	F4	F5
I	.062	.304	.059	.082	.277
I_G	.039	.190	.014	.044	.06
IGr	.052	.099	.014	.022	.04
CFS	No	Yes	No	No	No
CBF	Yes	Yes	Yes	Yes	Yes
PC	.738	.522	.393	.284	.188

B. Dataset Security

The current scheme further improves on the robustness level of earlier technique [7][8][9] by using the data-grouping strategy. For brevity, we compute the probability of successfully attacking one encrypted bit only. Let Prob be the probability of successfully attacking one row of the data having logical n groups. Since, the attacker is not aware of the secret parameters that were used during the data-grouping stage; therefore, he cannot intentionally target a particular group for launching different types of attacks. As a result, he will have to select a random feature to launch his attacks. In this case, the probability of successfully corrupting a encrypted bit is $(.5)^n$ because different encryption is inserted in different groups. Since, we are encrypting all rows in the dataand also using the majority voting as an error correction measure; therefore, the attacker has to target at least half the total number of rows to achieve his objective. If there are N rows in the dataset, the probability of successfully decrypting bit is$((.5)^{n)(N/2)})$. For large datasets and large values of , this probability becomes significantly small. Consider, for example, a data has 100 rows and n=4 . The probability of successful launching an attack on this data is:

$$Prob=((.5)^4)^{50}=(.0625)^{50}=6.22\times10^{-61}$$

This probability in earlier technique [7] is 8.88×10^{-16} therefore, the new technique has significantly reduced the probability of decrypting. It is important to emphasize here that the approaches like [5] and [6] do not cater for classification potential of the feature and do not provide. Moreover, [5] and [6] also require the data owner to define the data usability constraints. In comparison, the proposed usability constraints model can easily be integrated with any encryption decryption algorithm (including [5] and [6]) and does not require a data owner to manually define the usability constraints. Furthermore, our current approach is also significantly different from earlier approach [7] that also demands an owner of a dataset to manually define the usability constraints. The selected features could be only numeric features that have least information gain. In contrast, the current approach is applicable to all types of features and also provides a mechanism to encrypt high ranking features.

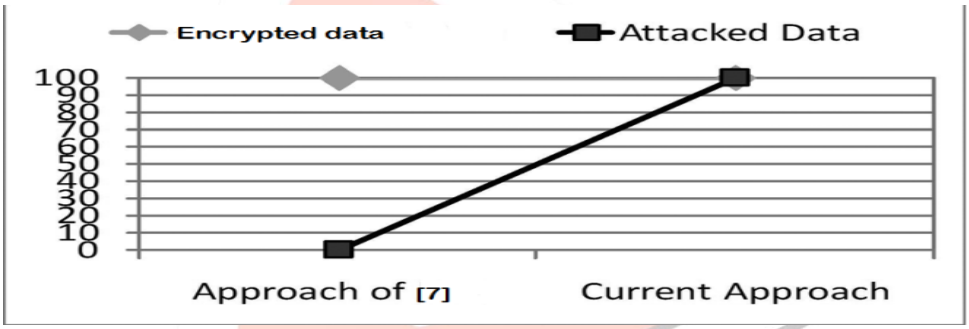

Fig 3 shows the difference in level of security when compared to earlier technique and proposed technique

VII. CONCLUSION

In this paper, a novel knowledge-preserving and lossless usability constraints model and a encryption/decryption scheme has been proposed for outsourcing datasets. The benefits of our techniques are: (1) identifying the vital characteristics of a dataset which need to be preserved during encryption; (2) ranking the features on the basis of their classification potentials; (3) logically grouping the data into different groups (clusters) based on this ranking for defining local usability constraints for each group; (4) defining global usability constraints for the complete dataset; (5) modeling the local and global usability constraints in such a manner so that the learning statistics of a classifiers are preserved. To the best of our knowledge, no technique in the literature exists that automatically computes "usability constraints" for a dataset that once enforced would preserve the knowledge contained in it. Moreover, the enhanced encrypting scheme can work with any type of data: numeric and nonnumeric with more 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 watermarking .

REFERENCES

[1] Kaggle's Contests: Crunching Numbers for Fame and Glory 2012[Online]. Available:http://www.businessweek.com/magazine/kaggles-contests-crunching-numbers-for-fame-and-glory-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] 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
[5] 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.
[6] R. Agrawal and J. Kiernan, "Watermarking relational databases," in *Proc. 28th Int. Conf. Very Large Data Bases*, 2002, pp. 155–166.

[7] 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.
[8] M. Kamran and M. Farooq, A Formal Usability Constraints Model for Watermarking of Outsourced Data Mining Datasets Tech. Rep. TR-59 Kamran, 2012
[9] M. Kamran and M. Farooq, A Formal Usability Constraints Model forWatermarking of Outsourced Data Mining Datasets . vol 8,no 6,june 2013
[10] RSA Algorithm A Short description Frans (frans@ic.vlsi.itb.ac.id)Department of Electrical Engineering Institut Teknologi Bandung Ganesha 10 Bandung, Indonesia