IOSR Journal of Engineering (IOSR JEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
PP 11-15

Comparative Study and Analysis of Various Privacy Preserving Data Mining Methods and Designing Efficient Method

Miss. Shraddha. M. Dolas[1], Prof. Mr. Y.M. Kurwade[2], Dr. V.M. Thakare[3]

Student in Department of Computer Science SGBAU, Amravati, Assistant professor in Department of Computer Science SGBAU, Amravati, Head of Department of Computer Science Department, SGBAU, Amravati

Abstract: *Privacy preserving plays vital role in designing various security related data mining applications. Protecting sensitive information in data mining has become an important issue. Data distortion or data perturbation is a critical component widely used to protect the sensitive data. This paper focused on five different techniques such as Visual crypto, Distributed SVM, Utility mining, ID3 mining over encrypted data, Pattern mining on event log data. But some problems are persisting in each method. The paper proposes the method to overcome the existing problems and the improved method "Privacy preserving decision tree learning using unrealized data sets" is proposed in this paper.*

Keywords: *Data mining, multiple keys, Encrypted data, Distributed system*

I. Introduction

The term predictive data mining is applied to identify the data mining concept with the goal or aim to identify the statistical or neural network model or the set of model that can be used to predict some response of interest. Predictive analysis is the use of data, mathematical algorithm and machine learning to identify the future events based on the basic system of data.The data mining is an analytical process designed to explore data in a search of consistent pattern and systematic relation between variable and then to validate the finding by applying the detected pattern to new subset of data. The ultimate goal of data mining is the prediction and the predictive data mining is the most common type of data mining and one that has the most direct business system. This paper, discusses five different privacy schemes such as Visual crypto [1], Distributed SVM [2], Utility mining [3], ID3 mining over encrypted data [4], Pattern mining on event log data [5]. But these methods also have some problem so to overcome such problems improve version of (mobility scheme) that is "Privacy preserving decision tree learning using unrealized data sets"

II. Background

Some studies on privacy predictive analytics models have been done to develop the privacy scheme in recent past years. Such schemes are:VC is a method of encrypting a secret image into shares such that stacking a sufficient number of shares reveals the secret image. It shares the binary images, usually presented in transparencies where each participant holds a transparency. Unlike conventional VC methods are not required obscure computation for recovering the secrets. [1]. The SVM is a supervised machine learning an algorithm based on the concept of an optimal separating hyper plane that can be used to solve classification problem. Many applications problems can be solve by using SVM such as image processing, computer vision, bioinformatics, and astrophysics [2]. In case of mining algorithm the application of high utility item set mining and privacy preserving utility mining cannot generate the high quality profitable item set according to the user specified mining utility threshold but also enable the capability of privacy preserving for a private secure information [3]. The emergence and development of Internet resulted in the generation of huge amounts of data, which are often distributed among different sites. Many organizations and companies attempted to mine the data .To process data mining, such as the ID3 algorithm over encrypted data without the cloud decrypting the data is very challenging task which is based on the weak computational power [4]. Big data promise a huge leap in analytics and data mining for industrial data sets and industrial companies is becoming aware of this potential in light of industries and the Industrial Internet. The concerns of privacy regarding potentially business critical data are a major hurdle for new successful analytics business [5]. This paper introduces five privacy scheme i.e., Visual crypto, Distributed SVM, Utility mining, ID3 mining over encrypted data, Pattern mining on event log data. The paper is organized as follows. **Section 1** Introduction. **Section 2** discusses Background. **Section 3** discusses previous work. **Section 4** discusses existing methodologies. **Section 5** discusses attributes and parameters and how these are affected on privacy models. **Section 6** proposed method and outcome result possible. **Section 7** concludes the outcome and possible results of paper. Finally **Section 8** includes the Conclusion.

III. Previous Work Done

In research literature, many privacy models have been studied to provide various privacy schemes and improve the performance in terms of preserving the data and improve their efficiency.

Akash Saxena et.al [1] has proposed peeling of partitions from original image, grouping the peeled portions in a random manner using schemes like modular process. This is analogous to long division process with successive distinct remainders providing the information to reconstruct original image.

Mohammed Z. Omer et.al [2] has proposed data that tacitly quite access to the data either at centralized or distributed form. Distributed data becomes a big challenge and cannot handle by a classical analytic tool. Cloud Computing can solve the issues of processing, storing, and analysing the data at distributing locations within the cloud.

Jerry Chun et.al [3] has proposed appropriate transaction for hiding sensitive high utility item sets from a database. The downward closure property and the pre-large concepts are adopted in the proposed algorithm to reduce the cost of rescanning databases. These concepts are also conducted to evaluate the performance of the proposed approach in execution time and the amount of side effects.

Xuan Wang et.al [4] has proposed internet resulted in the generation of huge amounts of data, which are often distributed among different sites. Many organizations and companies attempted to mine the data with cloud computing. However, given rise of various privacy issues, sensitive data need to be encrypted before outsourcing to the cloud. To process data mining, such as ID3algorithm, over encrypted data without the cloud decrypting the data is a very challenging task.

Alessandro Marrella et.al [5] has proposed event log data it is advantageous to achieve sophisticated analysis there exist several serious privacy issues in the paradigm. It also investigates through an industrial use-case the application of a framework for privacy preserving outsourcing of event-log data.

IV. Existing Methodologies

Many schemes have been implemented over the last several decades. There are different methodologies that are implemented for different models i.e., Visual crypto, Distributed SVM, Utility mining, ID3 mining over encrypted data, Pattern mining on event log data.

4.1: Visual crypto

VC scheme is a simple method to establish the mechanism of the system with the ability to prevent cheating. The VC is user friendly threshold with complementary cover images. The measure independent characterization of contrast optimal VC schemes enables to provide a characterization of optimal schemes to assess the contrast [1].

4.2 Distributed SVM:

There have been many privacy-preserving schemes designed for various classification schemes such as Support Vector Machine SVM is a powerful scheme to find an optimal solution to maximize the margin between the hyper planes and difficult points that close to decision boundary. SVM attracts much attention of researchers to focus on privacy-preserving. To training SVM need a kernel matrix which contains the value of every pair of data points the inner dot product that can used build the global SVM model without revealing the participates data [2].

4.3 Utility mining:

The knowledge discovery in database which also called as data mining, has become a powerful technique and commonly be used to discover interesting and useful knowledge from massive data. The downward closure property and the pre-large concepts are adopted in the proposed algorithm and because of this it reduce the cost of rescanning databases. These concepts are also conducted to evaluate the performance of the proposed approach in execution time and there are amount of side effects are created [3].

4.4 ID3 mining over encrypted data:

It constructs a decision tree in a top-down manner from a given set of samples. Classic ID3 decision tree algorithm is designed for a centralized database setting where raw data are stored in the central site for mining. The multiple parties with weak computational power need to run an ID3 on their databases jointly and outsource most of the computation of the protocol and databases to the cloud. To process data mining, such as ID3cover encrypted data without cloud decrypting the data is a very challenging task [4].

4.5 Pattern mining on event log data:

The outsourcing of pattern mining of an event log coming from the process industry. A common type of analytics for log files is frequent pattern mining. The goal is to extract relevant association rules, in order to predict the triggering of severe events in advance. It provides a way to anonymize event log data for pattern mining with no information loss and a strong privacy guarantee [5].

V. Analysis And Discussion

The visual cryptography shows that the parameters of an enhanced efficient half toning technique used in embedded extended VC strategy for effective processing. The VC measures independent characterization of contrast optimal to provide a characterization of optimal scheme to assess the contrast [1]. Distributed SVM improve collaborative efficiently of global mining models can construct for multiparty which having a large workload for collect data and can divide into small jobs which is better for protected and regulated to performs data mining tasks [2]. Utility mining scheme shows generation of valid combination of the particle of a structure which can gently reduce the computations of multiple database scan. It also compresses the valid combination of the item set in the database [3]. ID3 mining over encrypted data schemes measures the framework with the multiple keys that keys underlying the solution of several secured protocol that based on the efficient privacy preserving outsourced [4]. Pattern mining on event log data show that an outsourced data set satisfies not only support anonymity, but also explores set based attacks empirically. The concept of support anonymity is extended to privacy [5].

Table 1: Comparisons between different privacy multiple schemes.

Mobility scheme	**Advantages**	**Disadvantages**
Visual crypto	The time taken to handle slices in all partition is measured is less than the total time which is completely randomize the partition of slices and the advantage is that, this time is independent of the number of slices chosen within a partition.	If the first set can be in a sliced sub frame and this set can be in constructed foam in foam the next set is repeatedly reconstructed.
Distributed SVM	The multi parties are efficient global data mining model can construct, workload of data can divide into small jobs which is better for protected to perform data mining tasks.	Distributed data becomes a big challenge and cannot handle by a classical analytic tool and it can take lots of amount of time for preserving the data.
Utility mining	The pre large concept and the improved strategy concept in the mining concept in terms of running time so the task of the system complete in a while.	It only generates the valid combination of item sets existing in database which can avoid computational problem.
ID3 mining over encrypted data	It prevent reconstruction errors under independence noise, and analyse the security of scheme when collusions are occurs.	A more copy of same data does not require large privacy since the added noise may be filtered out.
Pattern mining on event log data	It provide the concept implementations of interesting event log data tasks to demonstrate the trade-off between privacy and utility.	Disadvantage of log data is that only separated or group foam of data is added to the released data.

VI. Proposed Methodology

The privacy preserving is a process of providing the security for sensitive data where analytics of predictive can create lots of value for organizations.

Privacy predictive scheme is important and it is not simple task to identify and discuss about various methods which is based on variety of components information security, robustness, and efficient way for different privacy preserving analytics models.

There are still problems which trouble in this field such as preserving the personal data about the user or the person because of lacks of data can be hacked of the user and lots amount of personal information of the user data. New privacy predictive analytics method called “Privacy preserving decision tree learning using unrealized data sets”.

Preservation of the model is for secured and preserve the data is propose here to overcome the problems of this scheme. As this scheme is depend upon the information or the data which is available on the system and also the current data which can be used by the user day by day.

In the proposed method for the analysis there is dataset can be used. For the person firstly login to the twitter dataset for processing the system. There is an component that cleaning of noise if there is an uncorrupted data or interrupt can occur then because of this dataset it can prevent the data also because of reduction and the replacement dataset it can secure the information by replacing a garbage data with the new data.

If this process can be failed then it gives the error message so because of that users can again visit to home page or can login it. Then if the data are in sustainable foam then the feature selection can be used. The main use of this selection is that it can create the simple model to make users easier and to avoid the curse of dimensionality. Then by applying secure virtual machine, logistic regression this classifiers can be used to enable multiple system instances to run on currently on a single system. By acquiring the result each result or final data can be store as well as upload the profile gallery of the person. Compare this result to generate the best file against the various parameters. Then because of various results generate the competitive chart or a graph for providing the security to each part. And then apply the privacy preserving method for secured the data.

Basic steps of algorithm:

Step1: The person firstly login to the twitter dataset for processing the system. There is component that cleaning of noise if there is an uncorrupted data or interrupt can occur then because of this dataset it can prevent the data.

Step2: If the data is cleaned then apply the reduction and the replacement dataset it can secure the information by replacing garbage data with the new data.

Step3: If this process can be failed then it gives the error message so because of that users can again visit to home page or can login it.

Step4: If the data is in complete foam. Apply the feature selection. The main use of this selection is that it can create the simple model to make users easier and to avoid the curse of dimensionality.

Step5: Compare this result to generate the best file against the various parameters. Then because of various results generate the competitive chart or a graph for providing the security to each part. And then apply the privacy preserving method for secured the data.

With the help of Privacy preserving decision tree learning using unrealized data sets concept the proposed method performed in small space and in less interval of time.

Diagrammatic representation of proposed method is shown as follows:

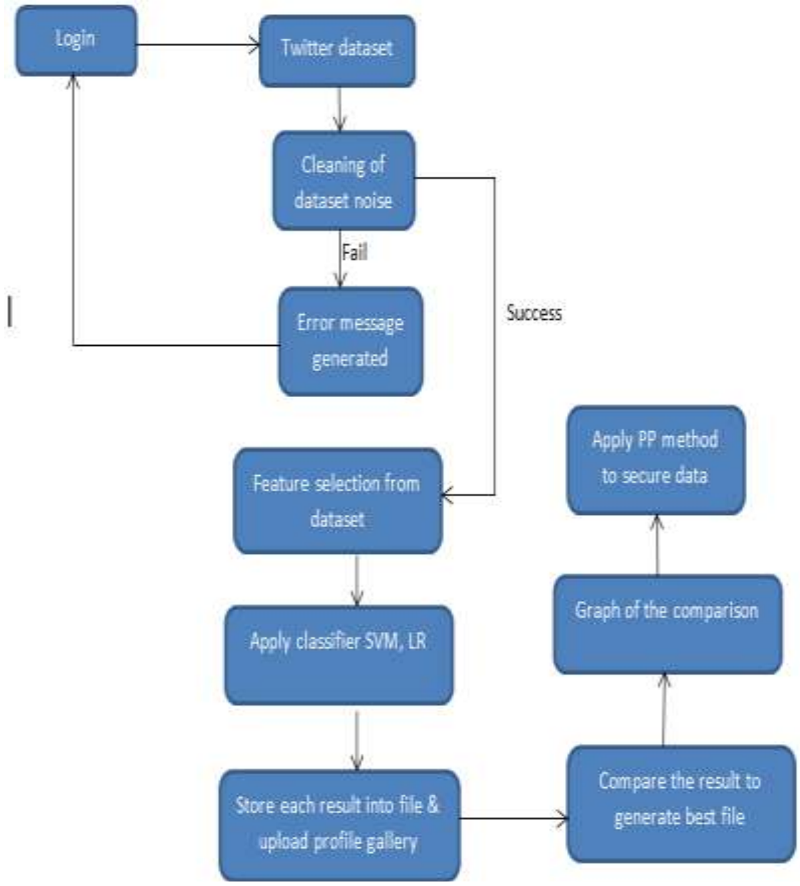
Login
Twitter dataset
Cleaning of dataset noise
Fail
Error message generated
Success
Feature selection from dataset
Apply classifier SVM, LR
Store each result into file & upload profile gallery
Compare the result to generate best file
Graph of the comparison
Apply PP method to secure data

Figure 1: *Analysis and design of privacy preserving with the data set*

VII. Outcome And Possible Result

The methods target to overcome the limitation of information loss while preserving the privacy. The collected data is used for various analysis or decision machining purpose by data mining. The preserving the data technique can prove to be useful and efficient in achieving the goal of gaining trust by preserving identify and privacy of the individual.

VIII. Conclusion

This paper focused on the study of various privacy preserving scheme i.e., Visual crypto, Distributed SVM, Utility mining , ID3 mining over encrypted data , Pattern mining on event log data. But there is some privacy problems associated with knowledge discovery from data so to improve this “Privacy preserving decision tree learning using unrealized data sets” privacy method for data mining is proposed here. When database moves through the dataset then the propose method provide the location for movement of the query in less time.

IX. Future Scope

From observations of the proposed method the future work will conduct the protocol on others algorithms to a trade-off between an efficiently and the accuracy.

References

[1]. Akash Saxena “Visual crypto” IEEE CONFERRANCE ON DATA MINING NETWORK TRACKING, 2016.

[2]. Mohammed Z. Omer, Hui Gao “Distributed SVM data mining” International Conference on Soft Computing & Machine Intelligence, 2016.

[3]. Jerry Chun, WenshengGan “Privacy preserving utility mining” Science Direct, 2016.

[4]. Xuan Wang, Ye Li, Zoe L. Jiang “ID3 mining over encrypted data” IEEE International Conference on Embedded and Ubiquitous Computing, 2017.

[5]. Alessandro Marrella, Anna Monreale “Pattern mining on event log data” IEEE International Conference on Cloud Computing Technology and Science, 2016.