ISSN (ONLINE): 2395-695X
ISSN (PRINT): 2395-695X

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Volume 2, Special Issue 19, October 2016

Privacy Preserving In Patient Health Record Using Data Perturbation And Decision Tree

S. Mekala [1], Dr. S. Sathappan[2]
M.Phil Full Time Research Scholar[1], Associate Professor[2]
Department of Computer Science,
Erode Arts and Science College,
Erode, Tamilnadu, India,
mekalasubu@gmail.com[1], devisathappan@yahoo.co.in[2]

***ABSTRACT*--In this paper, we provide an accurate and efficient privacy preserving data mining techniques in Patient Health Record. This system is used to provide security in patient health record that contains sensitive information. In this system we use 14 attributes such as Pid, Age, Gender, Sugar, Drugs, BP, HIV, Smoking, Disease, Diet, Allergies, Height, Marital Status, Class. To ensure security in patient health record we use data perturbation and for classifying the disease into sensitive and non-sensitive disease we using J48 decision tree. The data perturbation technique is implemented in MATLAB and the accuracy results are analyzed by using Weka 3.8 tool. The experimental results show that the accuracy before and after data perturbation techniques is 100%.**

***Keywords*: Data Mining, Privacy Preserving, Data Perturbation, Decision Tree, Cryptography**

I. INTRODUCTION

"Data mining is the computational process of discovering patterns in large data sets". In short it is used to extract knowledge from analysing data in different perspective. This discovered knowledge is used in the application medical industry[1]. A major challenge in medical industry is to provide security in patients health record which contain some sensitive information. Now a day it is important to maintain confidentiality of personal details in records maintained by the various fields.

According to the definition, we use privacy preserving data mining techniques to provide the confidentiality in patient health record maintained in the medical fields. By using this technique we can provide the security for each individual personal detail such as their ID, Name, Marital Status, Diseases, etc. In this paper, we use data perturbation techniques which are one of the Privacy preserving data mining techniques to hide the Patient ID who is suffered with the sensitive disease which may be either Breast Cancer or HIV.

II. PATIENT HEALTH RECORD

In this world the patient health record is very important to maintain the details of the patient such as Patient ID, Patient Name, Address, Sugar, BP, Drugs, Marital Status, Smoking, HIV, Age, Gender, Diet, Allergies, Height and Type Of Diseases. This records contain some sensitive information which is to be maintained confidentially to protect the patient who is affected by the sensitive diseases Cancer and HIV.

III. LITERATURE SURVEY

Hillol Kargupta et.al. [20] has used Data Perturbation to preserve data privacy by adding random noise, and estimate the pattern accurately. They use Randomization-based Techniques to generate random matrices.

Thanveer Jahan et.al. [21] has used Data Perturbation with SSVD for analyzing the system and used to transform original dataset into distorted data set using Sparsified Singular Value Decomposition. Finally the SSVD is more successful than SVD in his system. .

Y.Lindell et.al. [19] has used Cryptographic Technique to encrypt the data and also used a proper tool set for algorithm of cryptography. This approach is useful for small database. The result shows that the author achieved the security through this cryptographic techniques.

Shweta Taneja et.al. [22] has proposed the Hybrid C-Tree Algorithm for Privacy Preserving Data Mining . ASCII code and special characters to encrypt the sensitive information and the result shows that the privacy of medical data is preserved .

Anvita Srivatsava et.al. [23] has proposed the Privacy Preserving Data Mining in Electronic Health Record. Here the auhtor used k-anonymity techniques to protect the sensitive information from the Electronic health record and finally they achieved good result.

IV. PROPOSED SYSTEM

Today, medical fields maintain the health record for each patient separately which contain huge amount of data[3]. The main objective of this research is to provide security and privacy of information stored in the record using the data mining techniques maintained by the hospitals. To develop this system we have chosen the data perturbation method which will replace the sensitive information with "0" in the copy of the database and cryptographic techniques is

implemented in the Netbeans used to provide security for the original database so it is viewed only by the authorized user. This technique will provide better security than any other techniques in privacy preserving data mining.

A. DATA SOURCE

The patient health record database consists of 1000 records. The dataset consist of 14 attributes such as Patient ID, Age, Gender, Marital Status, BP, Sugar, Allergies, Diet, HIV, Drugs, Smoking, Disease, Height and Class and their description is listed below.

Table I. Description of 14 attributes

S. No	Attribute	Description	Values
1.	Pid	Patient Identification number	Continuous values
2.	Age	Age in years	Continuous Values
3.	Gender	Male or Female	0=Male 1=Female
4.	Sugar	Sugar Result	0=No 1=Yes
5.	Drugs	Drugs Result	0=No 1=Yes
6.	BP	Blood Pressure Result	0=No 1=Yes
7.	Smoking	Smoking	0=No 1=Yes
8.	HIV	HIV Result	0=No 1=Yes
9.	Disease	Sensitive/Non Sensitive disease	0=Non-Sensitive Disease 1=Sensitive Disease
10.	Allergies	Allergies	0=No 1=Yes
11.	Diet	Diet	0=No 1=Yes
12.	Height	Heights in Centimeter	Continuous
13.	Marital Status	Married or UnMarried	0=UnMarried 1=Married
14.	Class	Class	0=No 1=Yes

V. WEKA TOOL

Weka is a collection of machine learning algorithm for data mining tasks. These algorithms may be applied directly using the default algorithm in the tool itself or we can call the algorithm using our own java code[17]. This tool is useful for Pre processing, Classification, Regression, Clustering, Association Rules and Visualization.

VI. MATLAB TOOL

MATLAB(matrix laboratory)is a multi-paradigm numerical computing environment and fourth generation programming language. A proprietary programming language developed by math works. MATLAB allows matrix manipulations, plotting of functions and data, implementation of algorithms, creation of user interfaces, and interfacing with programs written in other languages, including C, C++, Java, Fortran, etc. The Matlab supports object oriented programming including classes, inheritance, virtual dispatch, packages, pass-by-value semantics, and pass-by-reference semantics.

VII. NETBEAN IDE

NetBeans is a software development platform written in Java. The NetBeans Platform allows applications to be developed from a set of modular software components called modules. It is cross-platform and runs on Microsoft Windows, Mac OS X, Linux, Solaris and other platforms supporting a compatible JVM. The NetBeans IDE is primarily intended for development in Java, but also supports other languages, in particular PHP, C/C++ .

VIII. DATA MINING TECHNIQUES USED IN THIS STUDY

In this paper we used three techniques such as Decision Tree , Data perturbation techniques, and cryptographic techniques are used to provide security for the patient health record.

A. Decision Trees

Now a day the decision tree approach is most frequently used for the classification problem. By using this technique we build a tree first and then the dataset is applied to it. There are different types of decision tree are available such as CART, ID3, C4.5, CHAID, and J48. From these types we have chosen the J48 algorithm for our system. This algorithm uses pruning method for building a tree. It is an extension of ID3. In the WEKA data mining tool, J48 is an open source Java implementation of the C4.5 algorithm. Pruning is a technique in machine learning which reduce the tree size by removing the over fitted data. It also reduces the complexity of the final classifier.

B. Data Perturbation

The Data Perturbation is a popular technique in privacy preserving data mining. This technique is used to provide data quality and also protect the privacy of information. It is relatively easy and effective technique for protecting sensitive data from unauthorized use. This method is classified into two main categories such as Probability distribution and Fixed

data perturbation. From these we have chosen the fixed data perturbation method which replaces all the sensitive disease patients' id into "0" in the patient health record. By applying these details in the database of sensitive disease patient record is maintained confidentialy. This method is well suited for the numerical and categorical data. The data type handle by this techniques are: Character type, Boolean type, Classification type, and Integer. These techniques are also named as "Data Distortion" or "Data Noise" and it is well suited for both the central and the distributed database. This technique is implemented in our system using the Matlab tool.

C. Cryptographic Technique

This technique is used to hide the data that are distributed across multiple sites which is legally prohibited from sharing. This cryptographic technique is applied after the perturbation techniques to provide more security. So it can be viewed only by the authorized user. It is one of the famous technique in data modification. Here the perturbated data is encrypted and the decryption technique is applied back to the encrypted data to get back the perturbated data. The cryptographic technique is applied by using the java coding in Netbeans IDE.

IX. RESULTS

The patient health record consist of 1000 records. The total records are divided into 10 folds each consist of 100 records and each time it take 1 fold as training and other 9 folds as testing data. Here we use weka 3.8 tool for our experiment. Initially the dataset contained some missing values that are identified and replaced with the most appropriate value by using the ReplaceMissingValues filter in weka tool. This process is said to be Data Pre processing. After Pre processing the classification techniques contain the J48 decision tree which is used to classify the data such as sensitive and non-sensitive disease and then we apply data perturbation method to hide the sensitive disease patient from the patient health record. After this the cryptographic technique is applied on the perturbated data to provide high security. The accuracy of classification before and after the data perturbation is 100%.

This accuracy is measured using the confusion matrix in weka tool consist of two classes a = "Yes"(Sensitive Disease)and class b="No"(Non sensitive disease)

Table II. A Confusion Matrix

Class	a	b
a(sensitive disease)	TP	FN
b(non sensitive disease)	FP	TN

TP(True Positive): Number of records classified as true.

FN(False Negative):Number of records classified as false.

FP(False Positive):Number of records classified as true but actually they were false.

TN(True Negative):Number of records classified as false but actually they were true

Table III. A Confusion Matrix for the classification.

Class	A	b
A	773	0
B	0	227

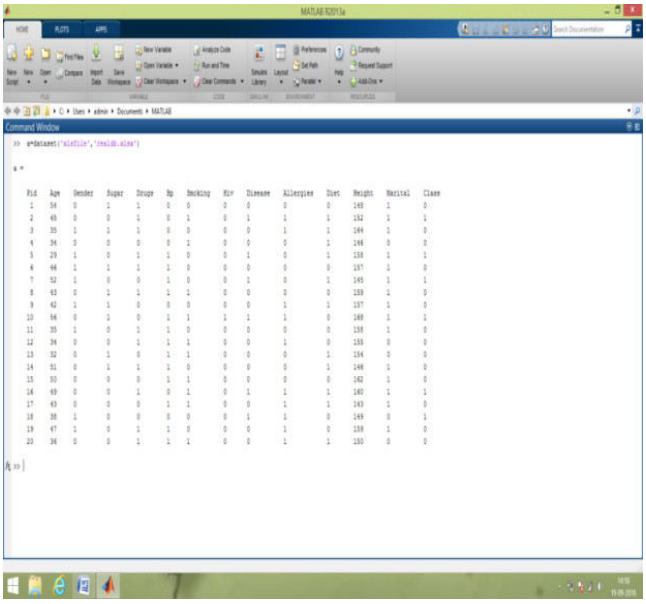

Figure 1. Snapshot of patient health record

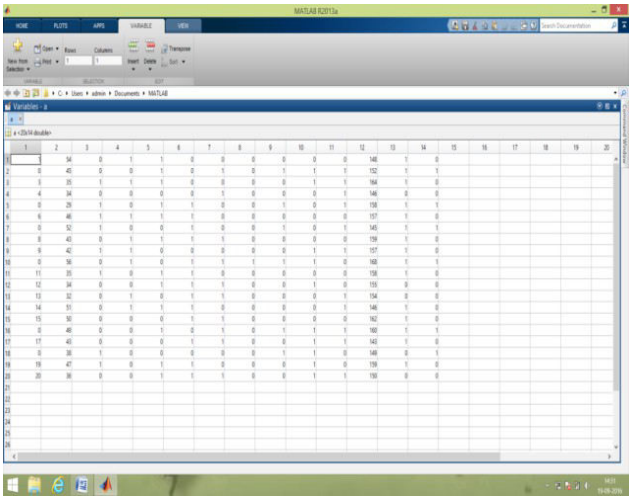

Figure 2. Snapshot of patient health record after applying data perturbation

In the above figure the data perturbation is applied to the patients who having sensitive disease such as cancer and HIV. The patient having sensitive disease means their patient

id is perturbated as “0”. So it is maintained confidentially. After this technique the cryptographic technique is applied to the database which will use the encryption and decryption method to protect the patient health record from the unauthorized user which will provide higher security than any other privacy preserving data mining techniques

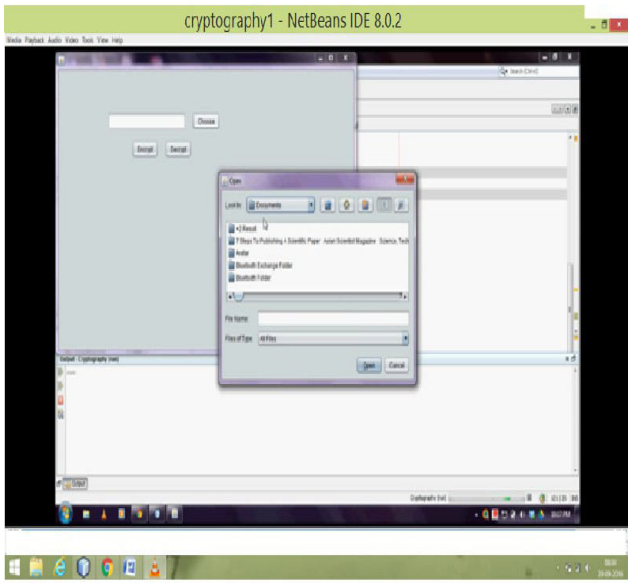

Figure 3. Snapshot of choosing file for applying cryptography technique

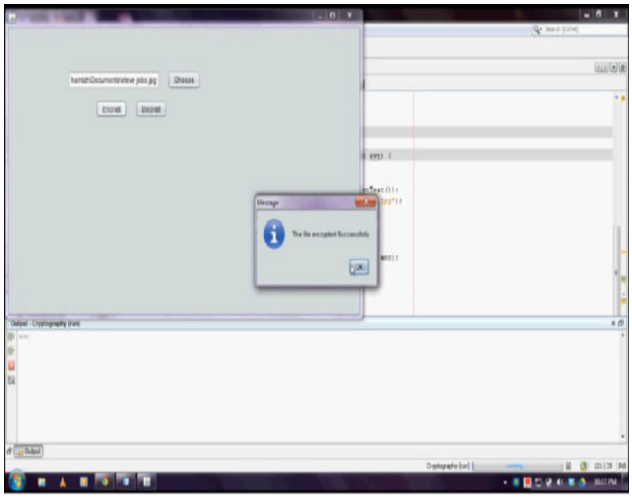

Figure 4. Snapshot of Encryption the database

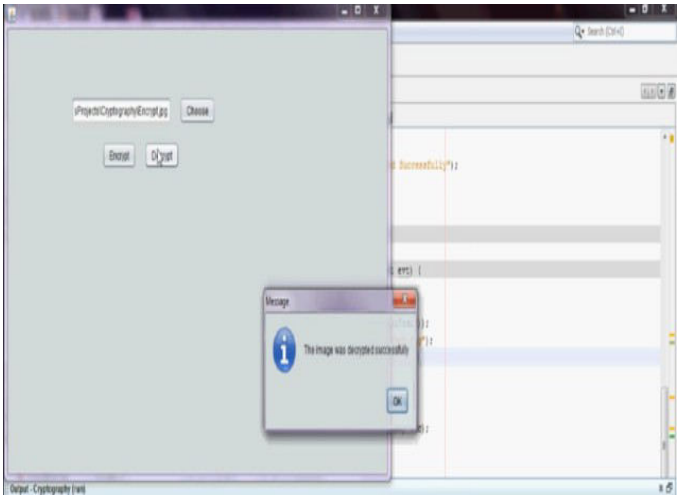

Figure 5. Snapshot of Decrypting the database

X. CONCLUSION

We have achieved privacy preserving in patient health record using data perturbation with the cryptography techniques which provide high security than any other security techniques in data mining. In our work, we have used WEKA tool for classification, MATLAB tool is used for data perturbation and fro cryptography the java coding is implemented in the Netbeans.

XI. REFERENCES

[1] Agrawal and Srikant, “Privacy Preserving Data mining”, Proceedings of the ACM SIGMOD International Conference on Management of data, 2000.

[2] Li Liu, Murat Kantarcioglu and Bhavani Thuraisingham, “The applicability of the perturbation based privacy preserving data mining for real-world data”, Data & Knowledge Engineering 65 (2008) 5–21.

[3] Samarati P, “Protecting respondent’s privacy in Microdata release”, IEEE Transactions on Knowledge and Data Engineering, 13:1010–1027.

[4] J. Han and M. Kamber , “Data Mining: Concepts and Techniques”, 2nd ed.,The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor 2006.

[5] M. B. Malik, M. A. Ghazi and R. Ali, “Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects”, in proceedings of Third International Conference on Computer and Communication Technology, IEEE 2012.

[6] P.Deivanai, J. Jesu Vedha Nayahi and V.Kavitha," A Hybrid Data Anonymization integrated with Suppression for Preserving Privacy in mining multi party data" in proceedings of International Conference on Recent Trends in Information Technology, IEEE 2011.

[7] R. Agrawal and A. Srikant, " Privacy-preserving data mining", in proceedings of SIGMOD00, pp. 439-450.

[8] T. Jahan, G.Narsimha and C.V Guru Rao, "Data Perturbation and Features Selection in Preserving Privacy" in proceedings of 978-1- 4673-1989-8/12, IEEE 2012.

[9] G. Nayak and S. Devi, "A Survey on Privacy Preserving Data Mining: Approaches and Techniques" in proceedings of International Journal of Engineering Science and Technology (IJEST), 2011

[10] S. Lohiya and L. Ragha, "Privacy Preserving in Data Mining Using Hybrid Approach", in proceedings of 2012 Fourth International Conference on Computational Intelligence and Communication Networks, IEEE 2012.

[11] A. Parmar, U. P. Rao, D. R. Patel, "Blocking based approach for classification Rule hiding to Preserve the Privacy in Database" , in proceedings of International Symposium on Computer Science and Society, IEEE 2011

[12] Jieh-Shan Yeh and Po-Chiang Hsu, "HHUIF and MSICF: Novel algorithms for privacy preserving utility mining," Jounal of Expert Systems with Applications, vol. 37, pp. 4779–4786, 2010.

[13] B. Pinkas. Cryptographic techniques for privacy-preserving data mining. SIGKDD Explorations, 4(2), Dec. 2002

[14] M. Prakash, G. Singaravel, "A New Model for Privacy Preserving Sensitive Data Mining", in proceedings of ICCCNT Coimbatore, India, IEEE 2012

[15] H. Kargupta and S. Datta, Q. Wang and K. Sivakumar, "On the Privacy Preserving Properties of Random Data Perturbation Techniques", in proceedings of the Third IEEE International Conference on Data Mining, IEEE 2003.

[16] Bharati M. Ramageri, Lecturer, Dept. of Computer Application, MIITR, Pune- Maharashtra, India, "Data mining techniques and applications", Indian Journal of Computer Science and Engineering, Vol 1 No 4 301-305 5.

[17] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An update; SIGKDD Explorations, Volume 11, Issue 1. [Available Online: http :// www.cs. waikato.ac.nz/ ml/weka/index.html]

[18] Martin Brown, "Data mining techniques", December 2012[Online]. Available: http:// www.ibm.com/ developerworks/library/ba-data-mining-techniques .

[19] *Y Lindell*, *B Pinkas*. Journal of cryptology 15 (3), 177-206, 2002. 1564, 2002. Introduction to modern cryptography. J Katz, *Y Lindell*. CRC PRESS, 2007.

[20] Hillol Kargupta, Souptik Datta, Qi Wang and Krishnamoorthy Sivakumar,2003 "Random-data perturbation techniques and privacy-preserving data mining" An International Journal Knowledge and Information Systems, ISSN: 0219-1377

[21] Thanveer Jahan, Dr. G.Narsimha and Dr. C.V Guru Rao[15] 2012 Privacy Preserving Clustering on Distorted data IOSR Journal of Computer Engineering (IOSRJCE) ISSN: 2278-0661, ISBN: 2278-8727 Volume 5, Issue 2 (Sep-Oct. 2012)

[22] Shweta Taneja, Shashank Khanna, Sugandha Tilwalia, Ankita, "A Hybrid C- Tree Algorithm for Privacy Preserving Data Mining", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-4, Issue-ICCIN-2K14, March 2014.

[23] Anvita Srivastava, Gaurav Srivastava "Privacy Preserving Data Mining in Electronic Health Record using K- anonymity and Decision Tree" International Journal of Computer Science & Engineering Technology (IJCSET)