ISSN 22773061

Privacy Preserving Data Mining using Attribute Encryption and Data Perturbation

Meenakshi Vishnoi[1], Seeja K.R[2]

Department of Computer Science

Jamia Hamdard University
New Delhi, India

[1]meenakshi.vishnoi85@gmail.com

[2]seeja@jamiahamdard.ac.in

ABSTRACT

Data mining is a very active research area that deals with the extraction of knowledge from very large databases. Data mining has made knowledge extraction and decision making easy. The extracted knowledge could reveal the personal information , if the data contains various private and sensitive attributes about an individual. This poses a threat to the personal information as there is a possibility of misusing the information behind the scenes without the knowledge of the individual. So, privacy becomes a great concern for the data owners and the organizations as none of the organizations would like to share their data. To solve this problem Privacy Preserving Data Mining technique have emerged and also solved problems of various domains as it provides the benefit of data mining without compromising the privacy of an individual. This paper proposes a privacy preserving data mining technique the uses randomized perturbation and cryptographic technique. The performance evaluation of the proposed technique shows the same result with the modified data and the original data.

Keywords

Data mining, Privacy Preserving, Data Perturbation, Cryptography.

Council for Innovative Research

Peer Review Research Publishing System

Journal: INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY

Vol 6, No 3

editor@cirworld.com
www.cirworld.com, member.cirworld.com

ISSN 22773061

1. INTRODUCTION

Data mining tool has made the process of extraction of knowledge and information very easy. Most of the organizations now [1][2] depend on data mining results for providing better services, achieving greater profit, and better decision-making. For these purposes organizations collect huge amount of data [3]. For example, to achieve the big profit and to apply the best business strategies business organizations [4] collect data about the consumers for marketing purposes, to provide the best treatment and medical research medical organizations collect medical records. As data mining includes large database which may consist of some sensitive and private information which can lead to the information loss behind the scenes of the data owner and this is one of the major problem. So, none of the organizations and data owner would like to share the data due to data loss. To solve this problem the concept of Privacy Preserving data mining (PPDM) has emerged as an active area of research which preserves both private and sensitive data. PPDM has provided the benefit of data sharing without the misuse of data and also provides the confidentiality of the data and data mining result. Various techniques such as randomization, perturbation, anonymization, swapping, etc are available for PPDM. If only perturbation is done on the data for privacy purpose, it could lead to the information loss for very critical dataset [5].Thus; we need to work towards minimizing both privacy loss and information loss. We have proposed here an approach of perturbation technique combined with the cryptography. In perturbation some noise has been added to the original data is modified the data and then the attribute name has been encrypted in such a manner that the result obtained from original data and the modified data remains the same providing the better accuracy.

2. RELATED WORK

Various techniques have been developed over past few years for PPDM.The data transformation based approach modifies sensitive data in such a way that it loses its sensitive meaning .In this process statistical properties of interest can be retained but exact values cannot be determined during the mining process[3]. Various data modification techniques such as noise addition [1] it applies random noise to the data sets without considering different privacy requirements of the different Users. One interesting reconstruction approach is proposed by Kargupta et al. in [3]. By using random matrix properties, Kargupta et al. [8] was able to reconstruct the approximate original data by separating the random noise from it. Data shuffling [7] technique which was used to maintain the confidentiality of the data but retains the analytical value of the confidential data. Data transformation [3] in which data is transformed preserving the privacy of sensitive attribute. Role based access of data [5] where each user has control access to data depending on their role. Data masking [6] technique provides protection of sensitive data and hiding of confidential data by modifying the sensitive data to create life-like false values. In Cryptographic techniques [3] the data is encrypted with encryption methods and a set of protocols are used to allow the data mining operation.

2.1 Randomization perturbation technique

In randomization perturbation approach the privacy of the data can be protected by perturbing [9] sensitive data with randomization algorithms before releasing it to the data miner. The perturbed data version is send to the miner to mine the patterns. The algorithm is chosen in such a manner that sensitive data is modified and it remains no longer a sensitive data and preserving the confidentiality. The original data is distorted through adding the noise component to the data which is obtained through randomization. Here each individual entry is added with the noise component. In a set of data records denoted by X = {x1 . . . x N }. For each record x_i a noise component is added which is drawn from the randomization method. The noise components are y1…yN. The distorted records are x1+y1….xN+yN.So, this new record is denoted as z1…zN. This method is very simple and also provides the confidentiality of data.

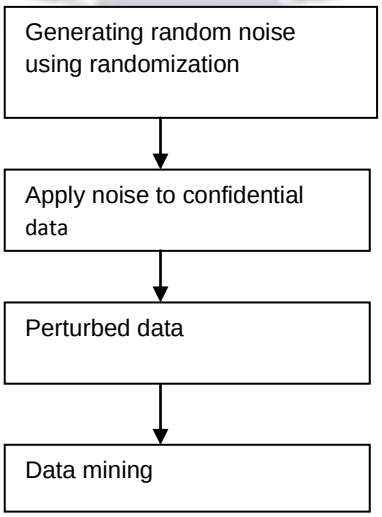

Fig1: Framework for perturbation technique

2.2 Cryptographic technique

Cryptography, the science of communication and computing in the presence of a malicious [3] adversary extends from the traditional tasks of encryption and authentication to protocols for securely distributing computations among a group of mutually distrusting parties.Cryptographic technique is used to encrypt the data to provide the security of data and hiding the individual information. In this paper we have used symmetric algorithm. This algorithm is faster and more secure. A DES algorithm has been used as a symmetric algorithm. We have used this algorithm to encrypt the attribute name of the dataset for preserving the privacy. The key generated is kept with the owner. The data miner has to work on the perturbed data and the encrypted attribute.

3. PROPOSED APPROACH

In this model we have generated a random noise using randomization method. This noise is then added to the original data. The data is modified in such a way that it becomes hard for third party to guess the original data. The field names of the dataset are encrypted and the data miner works on the perturbed and encrypted data.

3.1 ALGORITHM

Step 1

A random noise y_i is generated using randomized method. The original data is distorted using the noise component yi.

Step 2

This noise component generated in step 1 is added to the original data. Once the noise component yi is added the data becomes z1=x1 + y1

Step 3

Using cryptographic technique all attributes names is encrypted. Symmetric key encryption technique i.e. DES algorithm has been used to encrypt the attributes name.

Step 4

The key generated during encryption process remains with the sender and is not received by the miner

Step 5

The perturbed and encrypted data is then send to the data miner where the miner applies the mining algorithm. It can be any mining algorithm.

Step 6

The result generated is then send back to the owner in the encrypted format.

Step 7

Data owner after receiving the result decrypts the data using the key which was generated earlier during encryption process and removes the noise.

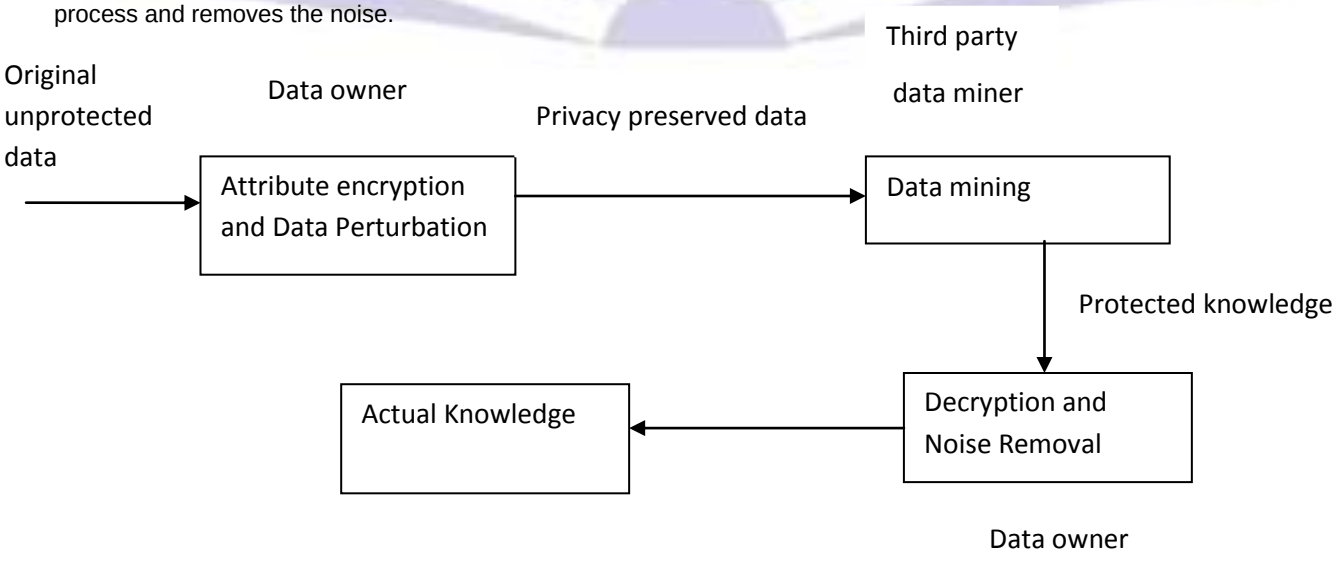

Fig 2: Framework for sharing the data using encryption and perturbation technique

3.2 Example

Let's take an example of a dataset Weather which is taken from UCI repository. This table contains 5 attribute and 14 records as shown in table 1. The dataset has been modified according to the proposed approach as follows.

Table 1: Original dataset of Weather taken from UCI repository

Outlook	Temperature	Humidity	Windy	Play
Sunny	85	85	False	No
Sunny	80	90	True	No
Overcast	83	86	False	Yes
Rainy	70	96	False	Yes
Rainy	68	80	False	Yes
Rainy	65	70	True	No
Overcast	64	65	True	Yes
Sunny	72	95	False	No
Sunny	69	70	False	Yes
Rainy	75	80	False	Yes
Sunny	75	70	True	Yes
Overcast	72	90	True	Yes
Overcast	81	75	False	Yes
Rainy	71	91	True	No

The above dataset will be modified according to the proposed approach.

1) The noise is generated using randomization method and original data is perturbed with the generated noise.
2) The attribute names are encrypted using DES.
3) The modified dataset is shown in table 2.

Table 2: perturbed and encrypted dataset

Xyfmmm	WQQulo	5b8gaQ	oOqwEv	FoHP/A
Sunny	110	110	False	No
Sunny	105	115	True	No
Overcast	108	111	False	Yes
Rainy	95	191	False	Yes
Rainy	93	105	False	Yes
Rainy	90	105	True	No
Overcast	89	90	True	Yes
Sunny	97	120	False	No
Sunny	94	95	False	Yes

Rainy	100	105	False	Yes
Sunny	100	95	True	Yes
Overcast	97	115	True	Yes
Overcast	106	100	False	Yes
Rainy	96	116	True	No

The above dataset is perturbed and encrypted dataset which is then send to the third party data miner who performs the data mining. The knowledge extracted is then send to the owner who then decrypts and removes noise from it.

It is found that the knowledge extracted from original data as well as Privacy Preserved data are found to be same as shown in table 3.

Table 3: Table displaying the result of both original and perturbed dataset.

Original Data	**No. of leaves**	**Size of tree**	**Correctly classified instances**	**Incorrect instances**
	5	8	9	5
Privacy Preserved Data	5	8	9	5
Knowledge Extracted fron original data Knowledge extracted from Privacy Preserved data are same				

4. RESULTS AND DISCUSSIONS

During this whole process data remains preserved at the owner site. The privacy is achieved because the data remains perturbed and encrypted during the mining process and key of encryption remains with the sender.Here we have taken a diabetes dataset from UCI repository. The dataset consist of 768 instances and 9 attributes.All the data except the class values are perturbed and all of the attribute names are encrypted using DES algorithm which has been implemented in java and data perturbation is done using randomized method. In this approach Weka tool has been used to obtain the result. Two types of result are displayed.

i) Pre-processing of attributes.
ii) Decision tree generated from J48 classifier for both perturbed and original dataset.

ISSN 22773061

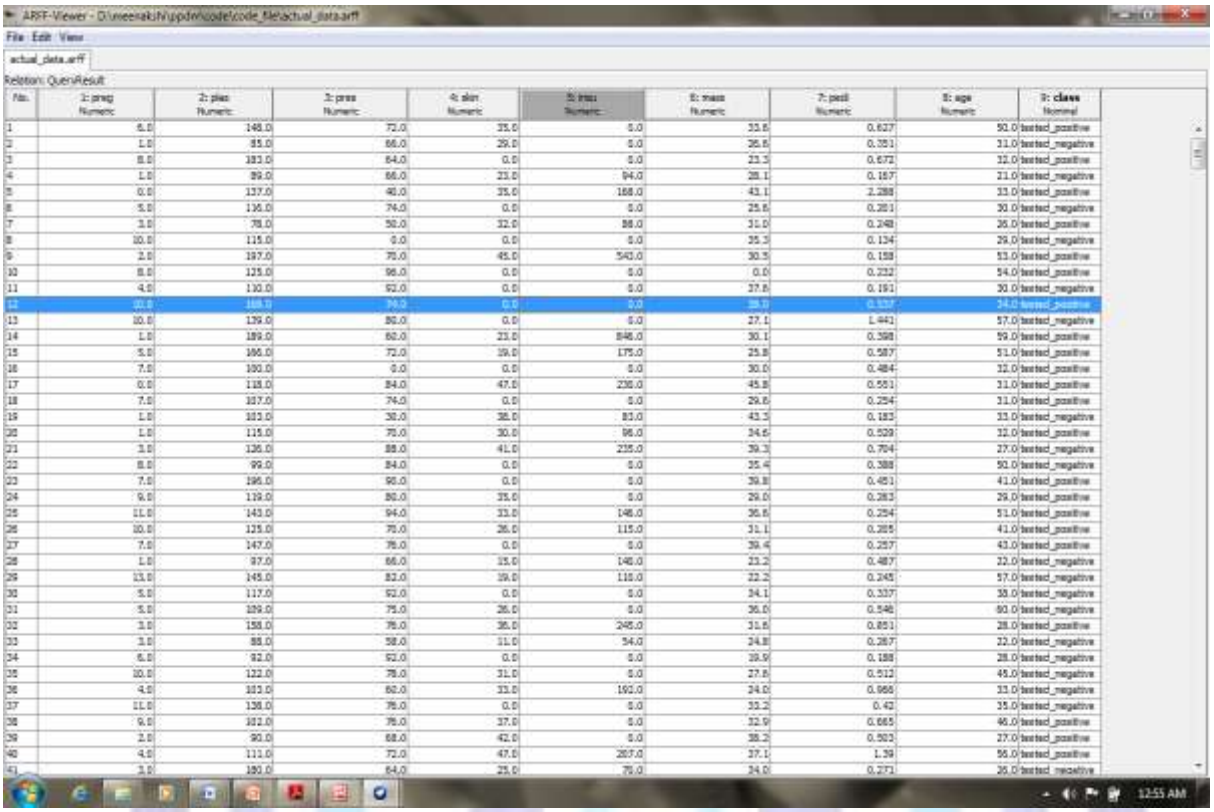

Fig 3: original dataset of diabetes taken from UCI repository

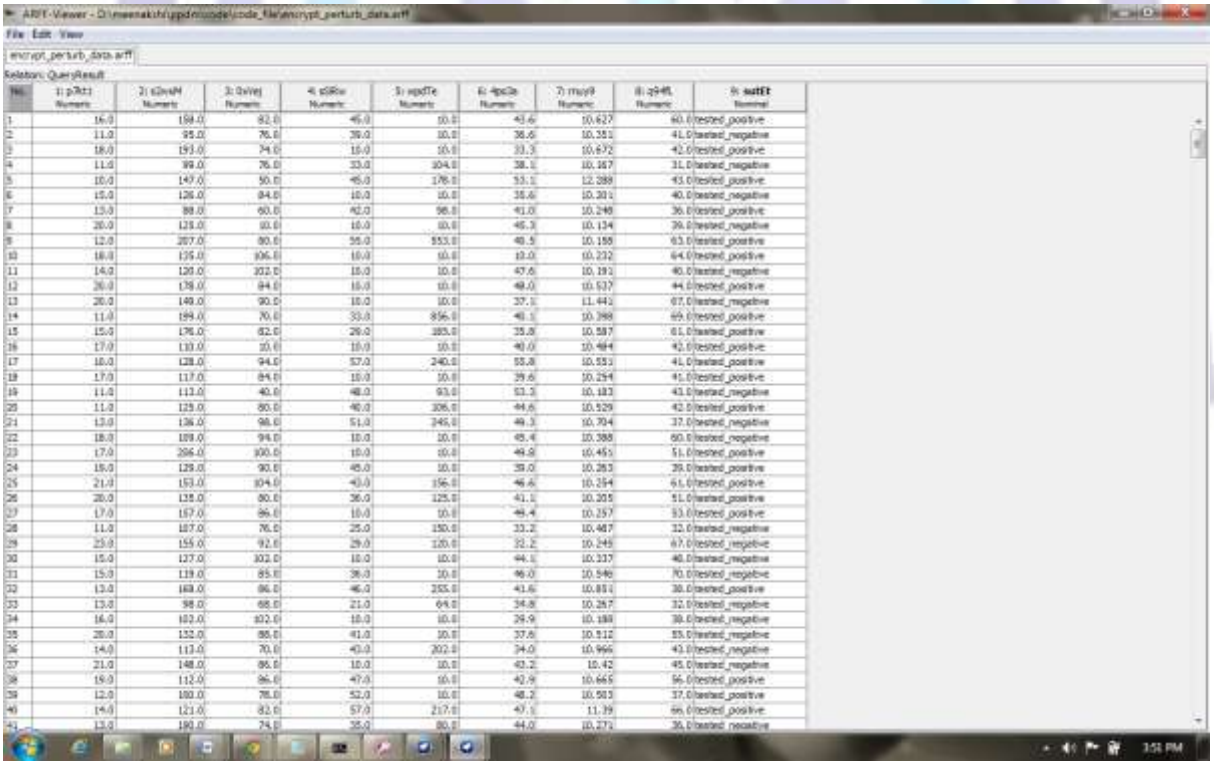

Fig 4: Perturbed and encrypted dataset

Below are the statistics which shows all the information of the values stored in each attribute of the dataset. The statistics differ according to the data type of the attributes. If the attribute is nominal, the list consists of each possible value for the attribute along with the number of instances that have that value. If the attribute is numeric, the list gives four statistics

describing the distribution of values in the data—the minimum, maximum, mean and standard deviation. The color coded statistics is displayed only if the attribute is nominal.There are two class values ‘tested-positive’ and’ tested-negative’. The red color is for ‘tested-negative’ class and blue is ‘tested-positive’ class. According to the range of values for each attribute the class values differ. For example for attribute ‘preg’ when the value is 0 or 1 the class value for tested_positive and tested_negative is 246 .

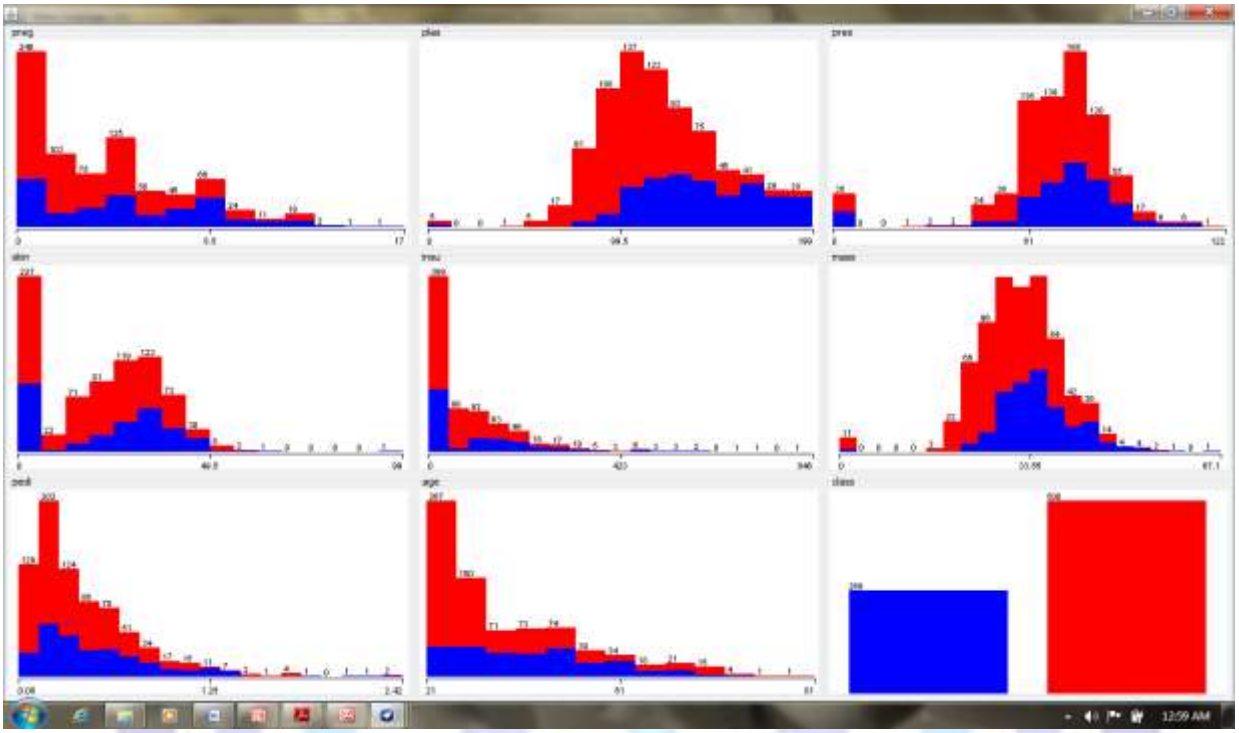

Fig 5: Result of original diabetes dataset

Below are the statistics which shows all the information of the values stored in each attribute of the dataset the values are perturbed and attribute names are encrypted the statistics generated is similar to the statistics of original dataset.

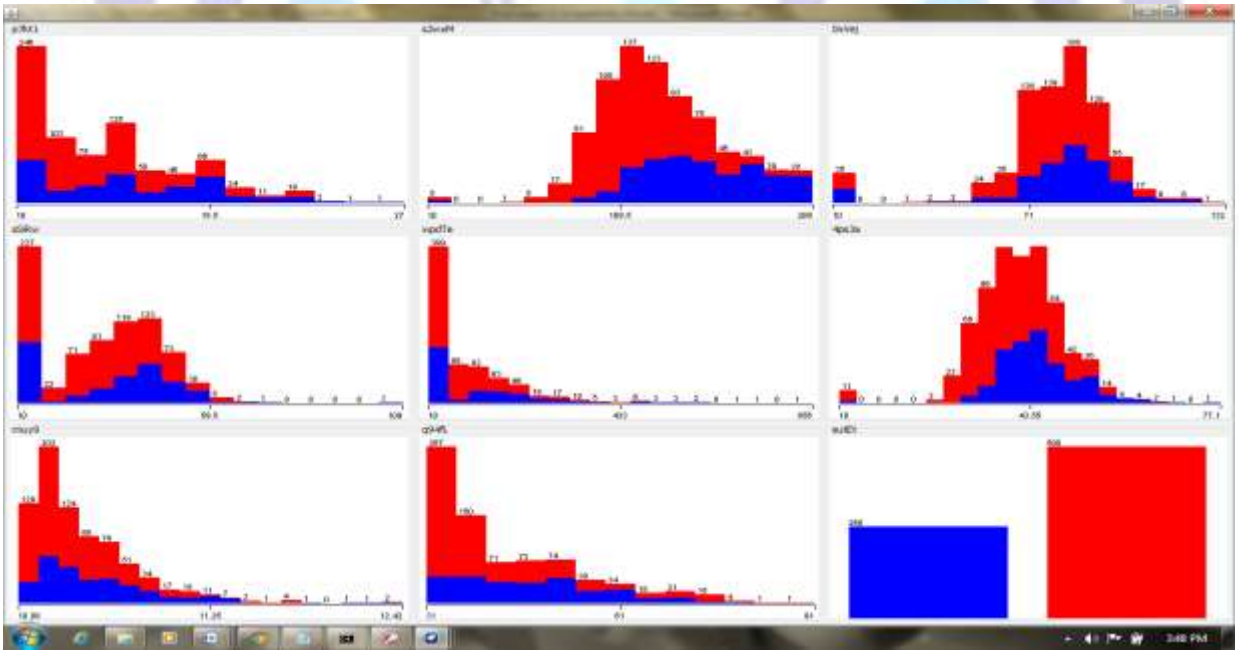

Fig 6: Result of perturb data

Below the table shows the result generated when actual and perturbed data are classified using J48 classifier.

Table 3: Table displaying the result of both original and perturbed dataset.

Actual Data	No. of leaves	Size of tree	Correctly classified instances	Incorrect instances
	20	39	567	201
Perturbed and Encrypted Data	20	39	567	201

Below is the tree generated after classifying the dataset in J48 classifier using Weka tool. The tree generated is a pruned tree .In J48 classifier. Here leaf nodes indicate which class an instance will be assigned to and should that node be reached. We can also see that the numbers are written in brackets after the leaf nodes indicate the number of instances assigned to that node, followed by how many of those instances that are incorrectly classified as a result.

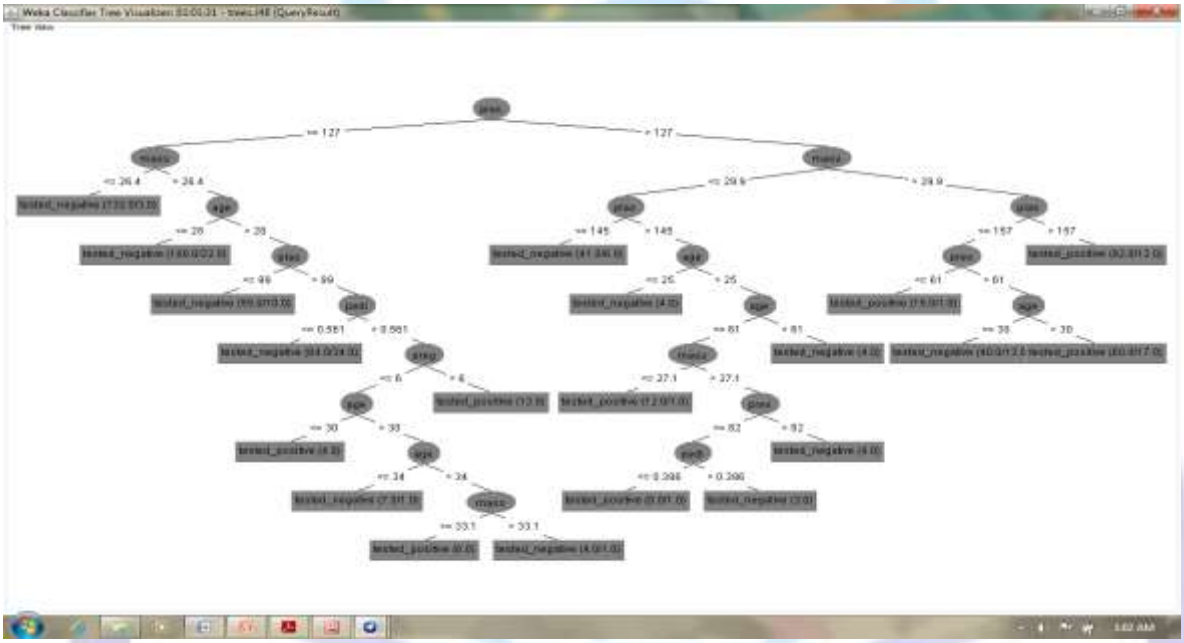

Fig 7: Decision tree generated from original data

Below is the tree generated after classifying the perturbed dataset in J48 classifier using the Weka tool. Here the nodes are the encrypted attribute name. And the tree generated is same as that of the tree generated using actual dataset.

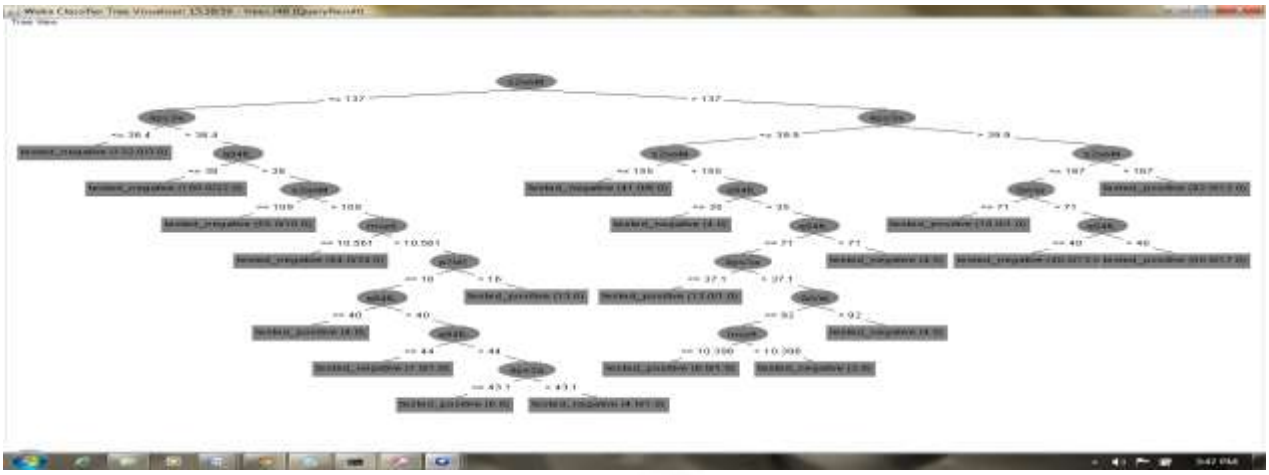

Fig 8: Decision tree of perturbed and encrypted dataset

5. CONCLUSION

Privacy preserving data mining is an emerging subfield of data mining. With the help of Privacy Preserving Data Mining sensitive data like medical data, banking data etc. can be protected without revealing the actual data to the third party data miner. In this paper, a privacy preserved data mining technique using data perturbation and encryption is proposed. The proposed approach is evaluated by using the weather dataset available in UCI repository. It is found that the knowledge extracted from original data as well as the protected data are same. Therefore the proposed approach can be used for mining sensitive data by protecting the individual privacy.

REFERENCES

[1] R.Agarwal and R.Srikant, “Privacy preserving data mining”, In Procseedings of the 19th ACM SIGMOD conference on Management of Data ,Dallas,Texas,USA, May2000.

[2] J. Canny, “Collaborative filtering with privacy”. In IEEE Symposium on security and privacy , pages 45-57 Oakland, May 2002.

[3] P.Kamakshi , Dr.A.Vinaya Babu.” Preserving Privacy and Sharing the Data in Distributed Environment using Cryptographic Technique on Perturbed data”, In Journal of Computing , volume 2, April 2010.

[4] The privacy preserving properties of random data perturbation techniques”, In Proceedings of the 3rd IEEE International Conference on Data Mining, pages 99–106, Melbourne, Florida, November 19-22, 2003.

[5] *Lalanthika Vasudevan , S.E. Deepa Sukanya, N. Aarthi*,”* Privacy Preserving Data Mining Using Cryptographic Role Based Access Control Approach”, in the Proceedings of the International Multi Conference of Engineers and Computer Scientists 2008 Vol I.

[6] An Oracle white paper, July 2010

[7] Muralidhar K. and Sarathy R. “ Data Shuffling- a new masking approach for numerical data.” Management Science, forthcoming, 2006.

[8] H. Kargupta, S. Datta, Q. Wang, K. Sivakumar, On the privacy preserving properties of random data perturbation techniques, in:

ICDM, IEEE Computer Society, 2003, pp. 99–106.

[9] R.Agrawal, A.Evfimievski, R.Srikant, “Information sharing across private databases”, In Proc.of ACM SIGMOD, 2003.

[10] www.cs.waikato.ac.nz/ml/weka/

[11] Mohammad Ali Kadampur, Somayajulu D.V.L.N.,” A Noise Addition Scheme in Decision Tree for Privacy Preserving Data Mining”, JOURNAL OF COMPUTING, VOLUME 2, ISSUE 1, JANUARY 2010.

[12] Krishna Priya .J Geetha Mary. A,” Perturbation of String Values”, In IJCSIT Vol. 2 (3) , 2011