International Journal of Advanced Computer Technology (IJACT)

ISSN:2319-7900

PRIVACY PRESERVING DATA MINING IN HEALTH CARE APPLICATIONS

First A. Dr. D. Aruna Kumari, Ph.d, ; Second B. Ch.Mounika, Student, Department Of ECM, K L University, chittiprolumounika@gmail.com;
Third C. A. Sai Kavya, Student, Department Of ECM, K L University, akaveetikavya@gmail.com;
Fourth D. M. Anvesh Babu, Student, Department Of ECM, K L University, mrbl4.anvesh@gmail.com

Abstract

In recent years data analysis and protection are the two important parameters in any organization to improve the efficiency and privacy. In hospital management, privacy plays a crucial role. In order to hide the patient's related databases effective data mining algorithms are to be implemented. In this project we are going to create a database and perform different algorithms on it, so as to make sure that private data is preserved. Our project mainly deals with choosing the best algorithm for performing the privacy preservation on the databases mined. In our project we use a weak tool to implement randomization technique and analyze how efficiently the data is preserved.

Keywords: Randomization, Weka, Privacy.

1. Introduction

In this digital age, with the advent of hybrid technologies personal information is being used in many aspects for example online banking, subscription to any news letters, registration in any websites for further access, entering patient details to a hospital database. Large number of databases is generated daily. For analyzing this huge stream of data many organizations encourage Data mining in order to make good decisions. While mining the data there is chance that sensitive information is exposed. If the sensitive information is released it gives threat to the individual to avoid this various privacy methods are applied. For example consider a hospital database which contains an attribute "date of birth". This is represented as 11^{th}nov 1993 in general manner. This field is used to analyze which disease is frequently occurring at what age. However during the analyzing of hospital databases for patient's record the patient age should not be revealed. Here some algorithms are implemented and date of birth is manipulated as **93. Here by doing such process the accurate age of the patient is disclosed and the occurrence of the disease will also be predicted so that the health statics will be predicted in well manner. In this paper we are going to analyze different methods of privacy preserving data mining such as randomization, anonymization and also by using

WEKA tool we are going to apply randomization technique on the clinical dataset and analyze the results.

1.1 Privacy Preserving in Data Mining

Privacy preserving data mining is an area of data mining to protect sensitive information. It gives a new ray to the field of datamining without interpreting the underlying data. Classification of privacy preserving data mining is given below [5].

Data Hiding	Data Perturbation	Value Distortion	Additive Perturbation
			Multiplicative Perturbation
			Data Microaggregation
			Data Anonymization
			Data Swapping
			Other Randomization Techniques
		Probability Distribution	Sampling Method
			Analytical Method
	Secure Multi-Party Computation (SMC) / Cryptographic Protocols		
	Distributed Data Mining (DDM)		
Rule Hiding	Association Rule Hiding	Data Perturbation	
		Data Blocking	
	Classification Rule Hiding	Parsimonious Downgrading	

Table 1. Classification of privacy preserving data mining

2. Related Work

Privacy preserving data mining has many applications and to secure the individual privacy we have many methods like anonymization, data swapping, randomization.

2.1 Anonymization

In the model the many attributes in the dataset may be in-conjunction with the other values that are used to uniquely identify the records. For example if the fields like DOB and ZIPCODE are used to uniquely identify the records by removing the sensitive fields from the values. The main idea in this model Is to reduce granularity that is not to uniquely identify the Kth record from the (K-1) set of records. Generalization and suppression are two techniques implemented in anonymization method which are used to secure the individual data record. In generalization, original value is masked and displays the value within the interval. Suppression represents the sensitive values are masked by (*).

ISSN:2319-7900

These two illustrated below

AGE	WEIGHT	DISEASE
20	45	ENT
40	60	HEART DISEASE

Table 2. Before generalization

AGE	WEIGHT	DISEASE
[15,30]	45	ENT
[35,55]	60	HEART DISEASE

Table 3. After generalization

NAME	ZIP CODE	DATE OF BIRTH
SAI	522004	ENT
DEEPTHI	522009	HEART DISEASE

Table 4. Before suppression

NAME	ZIP CODE	DATE OF BIRTH
SAI	5220**	ENT
DEEPTHI	5220**	HEART DISEASE

Table 5. After suppresion

2.2 Data Swapping

Additive and multiplication of noise are not only the distortion techniques. Data swapping is another for value distortion. Data swaping is dependent on the values of the neighboring the data unlike randomization which is implemented on the independent data. In this technique values of the records are interchanged and original data is not revealed to the researchers and privacy is preserved. Drawback of data swapping is the accurate results are not shown.

2.3 Randomization

This algorithm is used over the centralized data mining and allows the secure to sensitive data of an individual. For example, in hospital databases it has the patients related values like name,age,address,disease,contact. To apply the randomisation on this data we used the WEKA tool. We add some randomised value to the attributes which we want and we get modified dataset.

3. Description Of Work

The purpose of privacy preserving data mining is to analyze the data and also to maintain the privacy of an individual data by modifying the original dataset value to a nearest neighboring values. This can be done through number of techniques like data hiding, blocking, data randomization. Randomization is one of the technique which is used to modify the data from the original dataset. It is a prevailing method in present privacy preserving data mining studies. It disguise the values of the records by adding noise to the original data. The below figure shows, random noise that is random number is added to the original dataset and the original data now changes to the randomized dataset. This randomized dataset modifies the sensitive information of the individual data like age, contact number, pincode, address. So that the information is not revealed and also researchers can do the research on the particular application area.

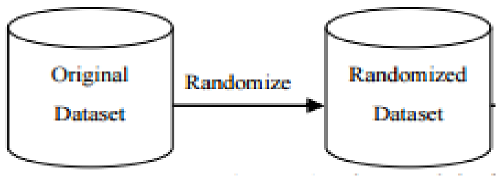

Fig 1. Model of randomization

The randomization method provides an effective yet simple way of preventing the user from learning precise data, which can be easily carried out at data collection phase to keep privacy data mining, because the noise added to a given record is independent of the behavior of other data records. When the randomization method is implemented, the data collection process consists of two steps. The first step is for the data providers to randomize their data and transmit the randomized data to the end user. In the second step, the end user estimates the original distribution of the data by employing a distribution reconstruction algorithm. In their randomization scheme, a random number is added to the value of a sensitive attribute. For instance, if ai is the value of a sensitive attribute, ai □ ri , rather than ai , will appear in the database, where ri is a random noise drawn from some distribution. It is displayed that given the distribution of irregular noises, regenerate the distribution of the original data is possible. The method of randomization can be illustrated as follows. Consider a set of data records denoted by A = {a1 . . . aN}. For record ai ! A, we add a noise component which is drawn from the probability distribution fB (b). These noise components are drawn independently, and are denoted b1 . . . bN. Thus, the new set of distorted records are denoted by a1 +b1 . . . aN +bN. We denote this new set of records by c1 . . . cN. In general, it is assumed that the variance of the added noise is vast enough, so that the original data record values cannot be easily guessed from the distorted data. Thus, the original records cannot be recovered, but the distribution of the original data records can be recovered. Thus, if A be the random variable denoting the data distribution for the original record, B be the random variable describing the noise distribution, and C be the random variable denoting the final data record, we have:

$$C = A + B$$
$$A = C - B$$

ISSN:2319-7900

Their are two kinds of perturbation are possible with the randomization method. Additive perturbation, randomized noise is added to the data records. The data distributions can be recovered from the randomized records. Multiplicative perturbation, the random projection of random rotation techniques, projection or random rotation techniques are used in order to perturb the records. This method includes random noise based perturbation and randomized response scheme. Hence it results efficiently and high information method.

4. Experimental Results

To implement the randomization technique we used the software called "WEKA"(Waikato Environment for Knowledge Analysis). This has the applications like Explorer, Experimenter, Knowledge flow, simple CLI. In explorer we follow steps like preprocess, classification, clustering, association, attribute selection, visualization. In Preprocessing the data is choosen and modified. In classification train and test learning schemes that classify or perform regression. In clustering the data is divided into clusters. In association we learn association rules for the data. Attributes selection, selects the most relevant attributes in the data. In visualization it views an interactive 2D plot of the data. This tool helps us to get the precise outputs. We take the example of the patient data related to diabetes which consists the attributes like preg, age, class. This sensitive information is modified to nearest neighbouring value which is not seen when the researchers want to research on particular area.

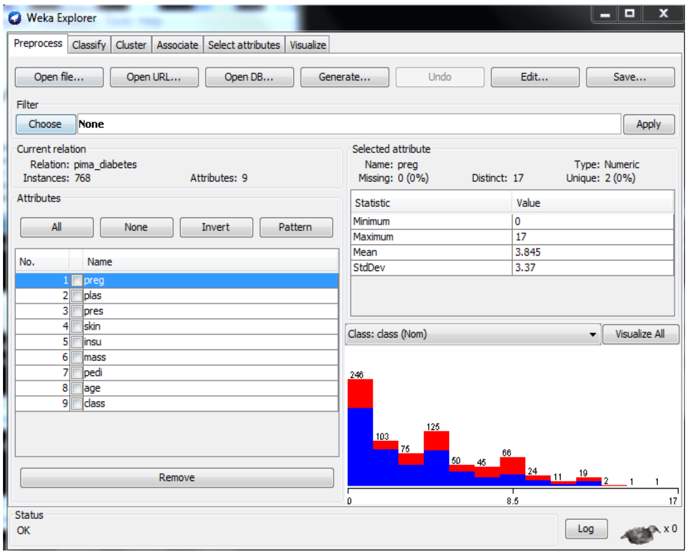

Fig 2. Original Dataset of patient related to diabetes

The figure(2) describes the original datset of the patient related to diabetes before the filtering is done. Bar graphs represents the number of people registered for that disease.

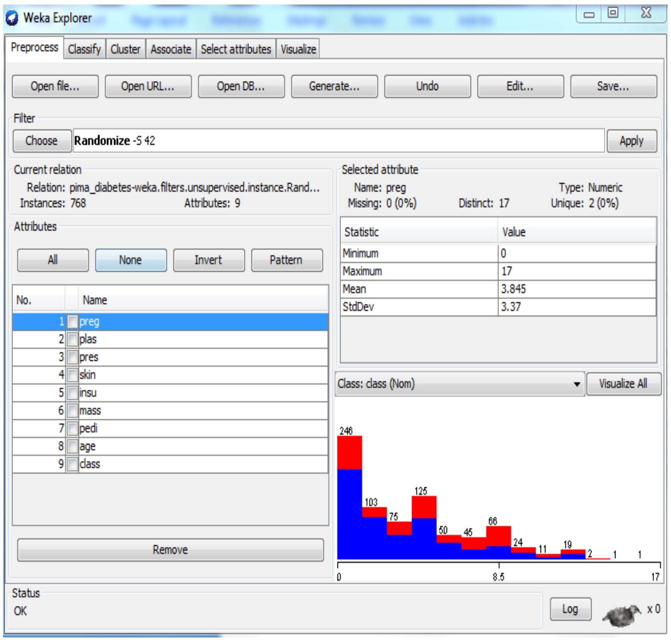

Fig 3. Preprocessing stage when randomization is applied

The figure(3) represents the preprocessed data after applying the randomization technique.

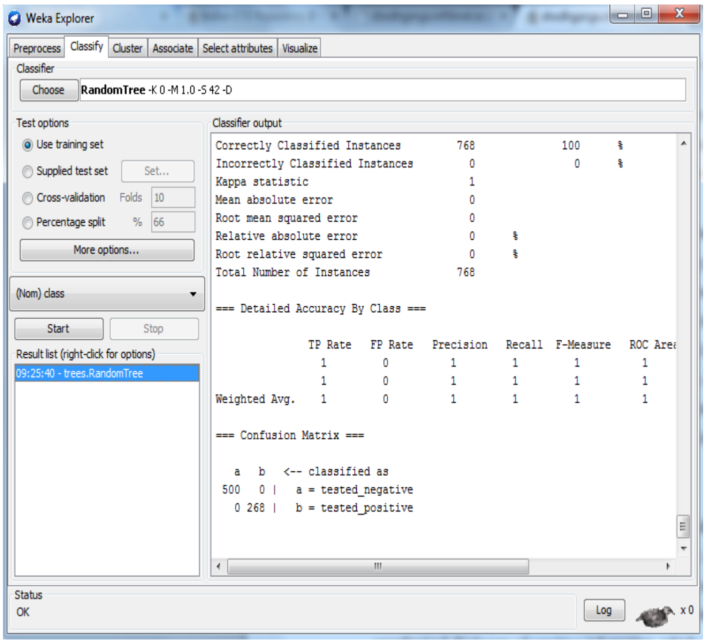

Fig 4. Classification of random tree classifier

The figure (4) represents the classification of data after random tree classifier is applied on the randomized data. The classifier output describes the correct instance for the particular class.

ISSN:2319-7900

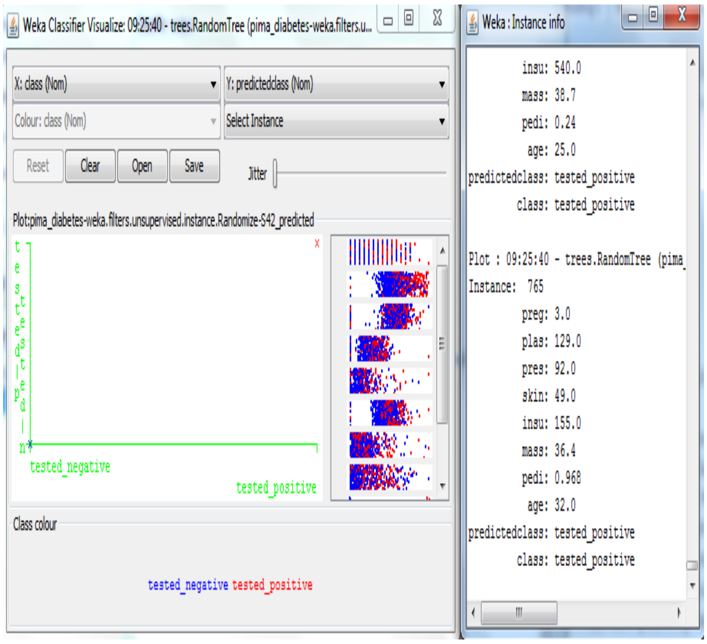

Fig 5. Classifier output

The figure(5) describes the classifier error output on the randomized data which we got from the preprocessing stage.

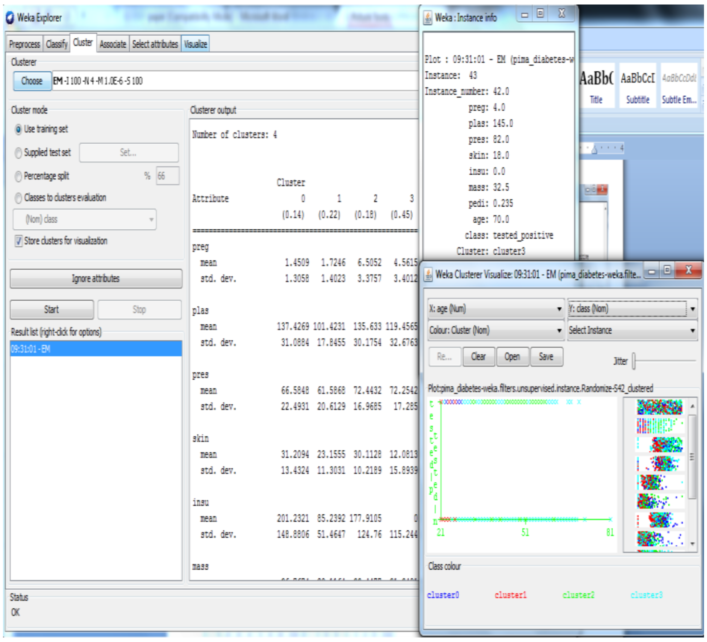

Fig 6. dividing the dataset into clusters

The figure(6) describes the clustering analysis of the randomized data.

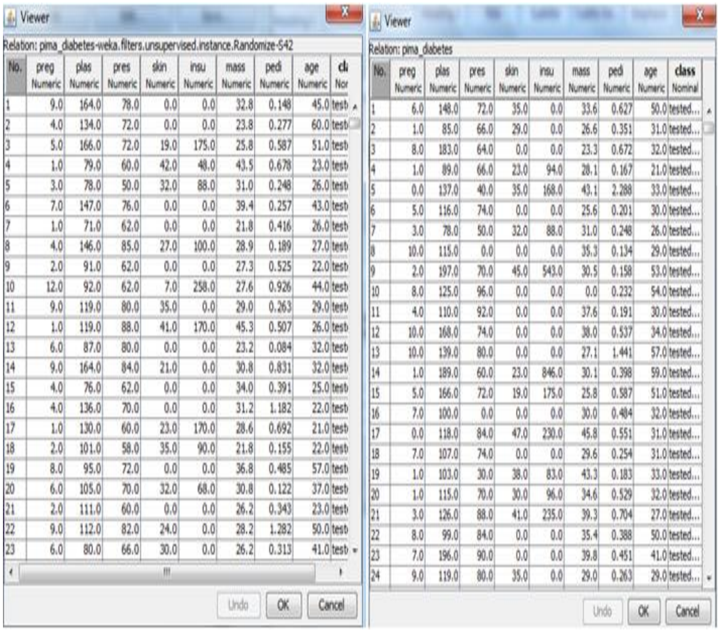

Fig 7. comparision of datasets

The figure(7) describes the comparision of original dataset and randomized dataset. We can observe that data is modified and the original dataset is secured.

5. Conclusion

From the results above stated it is clearly observed how the randomization algorithm is being implemented on the dataset that is taken and the results are properly analysed how much privacy is being done on the dataset by choosing the algorithm.

References

[1] N. Zhang, “Privacy-Preserving Data Mining”, Texas A&M University, pp.19-25, 2006.

[2] Z. Huang, W. Du, B. Chen, “Deriving Private Information from Randomized Data”, In Proceedings of the ACM SIGMOD Conference on Management of Data, Baltimore, Maryland, USA, pp.37-48, 2005.

[3] H. Kargupta, S. Datta, Q. Wang, K. Sivakumar, “On the Privacy Preserving Properties of Random Data Perturbation Techniques”, In Proceedings of the 3rd International Conference on Data Mining, pp.99-106, 2003.

[4] R. Agrawal, R. Srikant, “Privacy-Preserving Data Mining”, ACM SIGMOD Record, New York, vol.29, no.2, pp.439-450,2000.

[5] Methods and Techniques to Protect the Privacy Information in Privacy Preservation Data Mining N.Punitha R.Amsaveni.

[6] VECTOR QUANTIZATION FOR PRIVACY PRESERVING CLUSTERING IN DATA MINING D.Aruna Kumari , Dr.Rajasekhara Rao and M.Suman.

[7] Agrawal R., Srikant R. Privacy-Preserving Data Mining. Proceedings of the ACM SIGMOD Conference, 2000.

ISSN:2319-7900

[8] D.Aruna Kumari , Dr.K.Rajasekhara Rao,M.Suman "Privacy preserving data mining: LBG Algorithm" in International Journal of Database Management Systems(IJDMS) ISSN : 0975-5705 (Online); 0975-

5985(Print).
[9] D.Aruna Kumari, Dr. K.Rajesekhara Rao, M.Suman published a paper on "Vector quantization for privacy preserving clustering in data mining" in Advanced Computing: An International Journal (ACIJ -Nov 2012)
[10] D.Aruna Kumari, Dr. K.Rajesekhara Rao, M.Suman Tharun Maddu Published a paper on"Compression in privacy preserving data mining" in International Journal of Advanced Computer technology(IJACT ISSN : 2320-0790). April 2013.
[11] D.Aruna Kumari, Dr. K.Rajesekhara Rao and M.Suman Published a paper on "Privacy preserving clustering data mining using VQ code book generation" in AIRCCJ Computer Science and Information technology (CS &IT) Proceedings.

Biographies

FIRST A. Dr. D. Aruna Kumari, **B. Tech, M. Tech, Ph.d** received the Ph.D. degree from the K L University, Vaddeswaram, AndhraPradesh. Currently, She is an associate Professor of K L University. Her teaching and research areas include in data mining and published papers on privacy preserving in data mining.