DOI: http://dx.doi.org/10.26483/ijarcs.v9i2.5777

ISSN No. 0976-5697

Volume 9, No. 2, March-April 2018

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

PRIVACY PRESERVING USING ENSEMBLE CLASSIFICATION FOR HEART DISEASE DATA SETS

Anbarasi M.S.
Department of information technology
Pondicherry Engineering College
Puducherry, India

Shanmugapriya G.
Department of Information Technology
Pondicherry Engineering College
Puducherry, India

Silna K.
Department of Information Technology
Pondicherry Engineering College
Puducherry, India

Jyothish Kiran
Department of Information Technology
Pondicherry Engineering College
Puducherry, India

Krishna P.
Department of Information Technology
Pondicherry Engineering College
Puducherry, India

Abstract: In this modern era of thriving technology, the data being gathered through way of private in addition to public businesses is increasing each day. But in recent times people are more worried about their data and privateness being preserved at the same time as use of their data in other analysis purpose. Thus Privacy-Preserving Data Mining (PPDM) method has been proposed to permit the extraction of understanding from data at the same time as keeping the privateness of people. The primary purpose of our project is on preserving privacy for healthcare records as privateness lacks in Medical data. Privacy-Preserving Data Mining (PPDM) offers with shielding the privacy of individual's data or sensitive data without the utility of data. Therefore the strategies like anonymization, randomization are used to attain the intention. However, unfortunately anonymization results in certain level of information loss while preserving privacy. In order to overcome this problem, perturbation technique is carried out. Our challenge initiates with cleaning and preprocessing followed by ensemble classification and proceeded with perturbation to attain the goal. This method focuses on preserving privacy by perturbing the sensitive attributes in the Medical data without causing loss to the information in the process.

Keywords: Data mining, adaboost, perturbation technique, noise generation.

I. INTRODUCTION

Data mining is a technique employed to extract particular information from the available pool of statistics. But a some of this data may be non-public and confidential and for this reason need to be protected. This information that's categorized as sensitive attribute must not be exposed because it consists of non-public data of individuals. For this reason Privacy Preserving Data Mining (PPDM) techniques are employed to guard this information.

For this proposal, coronary heart disease dataset is selected for privacy preserving due to the fact cardiovascular disorder (cvd) is a primary purpose of dying. an expected 17.5 million people died from cvd in 2012, representing 31% of all global death[7]. Heart disease dataset has been taken from the uci web data repository which includes 14 attributes. The attributes are age, intercourse, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, and num.

The system includes 3 modules (i) In cleaning and preprocessing stage, the heart disease data set is in character form. Every individual may have distinct interpretation so as to avoid that interpretation we're converting individual data set into specific or numerical statistics set. So that the categorizing the data will be easy. Data is cleaned through processes such as filling in missing values the missing values are filled by correlating the record of same type and category of different patients.

(ii) In data mining, Classification is one of the most important method in data mining. The input to the classification method is data set having a number of attributes. The attributes values of the data set are either continuous or categorical. Ensemble classification are learning algorithms that construct a set of weak classifiers or learner after which classify new information by taking a weighted vote of their predictions. Maximum of the common classifiers in machine learning algorithm were used for heart disease diagnosis. It is now recognized that no single classifier can works best on all given problems. To overcome the drawbacks of single classifier, ensemble model is used. The method of combining classifiers is proposed as a new strategy to improve the overall performance of individual classifiers.

(iii) In this module sensitive attributes are decided on through attribute selection mechanism and those attributes are included by way of perturbation primarily based PPDM technique. Perturbation approach gives very low computation and communication value compared to

different PPDM techniques. In perturbation, the original values of an individual are replaced with some synthetic data or noise generated. So, that the statistical information obtained from the perturbated data does not vary from the statistical information obtained from the original data to a larger extent. Perturbation can be carried out by means of the usage of additive noise or multiplicative noise generation. Therefore this technique is greater convenient for maintaining privacy.

II. LITERATURE SURVEY

This section elaborates the earlier works based on the ensemble classification and Privacy Preserving Data Mining (PPDM) technique.

[1] This paper, aims to employ and evaluate such methods on learning analytics by approaching the problem from two perspectives: (1) the data is anonymized and then shared with a learning analytics expert, and (2) The learning analytics expert is given a privacy-preserving interface that governs her access to the data. We develop proof-of-concept implementations of privacy preserving learning analytics tasks using both perspectives and run them on real and synthetic datasets. We also present an experimental study on the trade-off between individuals' privacy and the accuracy of the learning analytics tasks. The application of state-of-the-art privacy-preserving data publishing and mining methods to learning analytics. They conclude by stating that technical solutions for privacy are most beneficial if there is a common demand from all parties, i.e., academics, practitioners, data and system owners, and students.

[2] In this paper, they suggested AdaBoost as classifier ensemble that can incorporate different base classifiers into classifier ensembles models for classification problems. This paper investigates the impact of using different base classifiers on classification accuracy of AdaBoost classifier ensemble. Classifier ensembles with five base classifier have used on five medical data sets. These results evaluated and compared choosing different type of decision tree algorithms for base classifier.

[3] In this paper they have elaborated the hybrid approach combining suppression and perturbation for Privacy Preserving data mining takes care of these requirements. This method focuses on the goal of preserving privacy by suppressing and perturbing the quasi identifiers in the data of online shopping customers stored on centralized data repository without causing any loss to the information in the process. The method targets to overcome the limitation of information loss while preserving privacy. The experiment is carried by setting up a local server on the system and the simulation results are compared with anonymization to show that the goal to achieve privacy of quasi identifiers without information loss is successfully achieved.

[4] The PPDM approach introduces Multilevel Trust (MLT) on data miners. Here different perturbed copies of the similar data are available to the data miner at different trust levels and may mingle these copies to jointly gather extra information about original data and release the data is called diversity attack.

To prevent this attack MLT-PPDM approach is used along with the addition of random Gaussian noise and the noise is properly correlated to the original data, so the data miners cannot get diversity gain in their combined reconstruction. Prevention of diversity attack can be done by appropriately correlating noise across at various trust levels and proved that the noise covariance matrix has corner wave property, and then the data miners have no diversity gain.

[5] In this paper, they concentrated on data perturbation procedures, i.e., Adding noise to the data in command to check thorough release of trusted values. The additive noise still permits the aggregate information to be read, about the overall collection of data but does not give away accurate values. The noise is a small randomly generated (or using certain algorithms), and added to the data. Hence, by this method we protect individual information and release information at the same time.

[6] This paper questions the utility of the random-value distortion technique in privacy preservation. The paper first notes that random matrices have predictable structures in the spectral domain and then it develops a random matrix-based spectral-filtering technique to retrieve original data from the dataset distorted by adding random values. The proposed method works by comparing the spectrum generated from the observed data with that of random matrices. This paper presents the theoretical foundation and extensive experimental results to demonstrate that, in many cases, random-data distortion preserves very little data privacy. The analytical framework presented in this paper also points out several possible avenues for the development of new privacy-preserving.

[7] In this study Neural network technique is adopted for classification of medical dataset. The experiment is conducted with Heart Disease dataset by considering the single and multilayer neural network modes. Back propagation algorithm with momentum and variable learning rate is used to train the networks. To analyze performance of the network various test data are given as input to the network. Parallelism is implemented at each neuron in all hidden and output layers to speed up the learning process. The experimental results proved that neural networks technique provides satisfactory results for the classification task.

III. PROPOSED PPECH SYSTEM

This concept initiates with the accurate value generation module by categorical conversion and pre-processing of heart disease data. As preprocessing is the most important stage in data mining process because real time data is often incomplete, inconsistent. Then followed by ensemble classification. Ensemble classification is a method of learning algorithms that construct a set of classifiers and then classify new data points by taking a weighted vote of their predictions. After finding the performance evaluation through ensemble model, additive data perturbation is carried out by noise generation. The noise is properly correlated to the original data to get the perturbated records. Before applying perturbation attribute selection will be

carried out in order to identify sensitive attributes from the original data. So the attacker cannot perform the recover sensitive information from the published data.

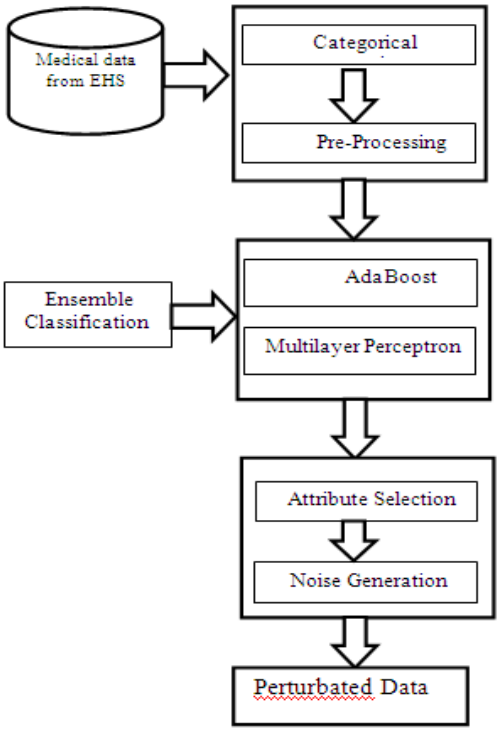

Figure 3.1: PPECH system Architecture

A. ***Categorical Conversion***

In this phase the medical information which has character values with distinct possibility of values are converted into categorical data or numerical facts. Due to the fact every individual could have distinctive interpretation. So with a purpose to avoid that interpretation we are converting character data set into categorical or numerical data set. So that the categorizing the data will be easy for further classification and prioritizing.

B. ***Pre-Preocessing***

Data pre-processing is an regularly unnoticed however an important step in the data mining procedure. Because the real time data will be often be incomplete and inconsistent of the mining outcomes raw data is pre-processed on the way to improve the efficiency of the mining system. The general drawback in weka tool for value generation is replacement of all missing values in a dataset using the mean of each attribute. To overcome the mean value generation the accurate value generation module is developed.

For accurate value generation the missing values are filled by correlating the record of same type and category of other patients. Initially the data has been classified primarily based on gender then the record with missing value may be compared with different patient's record using Euclidean distance and the distance with the lower value will be filled in the missing value. So, that the entered value will be more accurate than mean of each attribute.

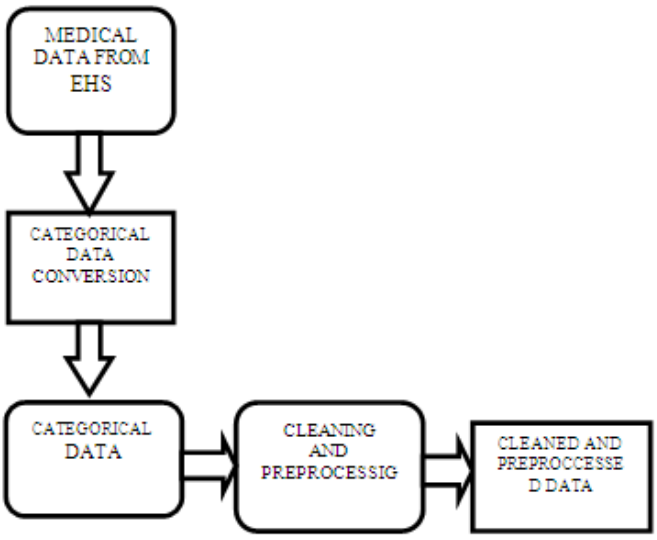

Figure 3.2: Accurate value generation

C. ***Ensemble Classification***

Ensemble methods use a combination of classifiers to increase accuracy and the performance of individual classifiers. An ensemble classifiers ought to be more correct than any of its individual or single classifier. Most of the common classifiers from the machine learning community had been used for heart disease diagnosis. It's now identified that no single model exists that is superior for all problem, and to overcome the limitation of single classifier the concept of ensemble model is introduced. The concept of combining classifiers is proposed as a new direction for the improving the performance of individual classifiers. Ensembling of adaboost and multilayer layer perceptron neural networks classifiers significantly will increase the sensitivity but decreases the specificity.

D. ***Adaboost***

Adaboost, brief for adaptive boosting, is an algorithm that offers a strong classification by way of combining numerous weak classifiers resulting into strong classifier. In this algorithm Naive Bayes and j48 are used as weak classifiers. So that Final learner could have minimum error and maximum learning rate resulting to the high accuracy. Adaboost is sensitive to noisy facts and outliers.

Advantage of using AdaBoost to improve the performance:

1. It produces a final classifier whose error rate or misclassification rate can be decreased by combining many classifiers whose error rate may be high.
2. Gives a strong classifier whose variance is significantly lower than the variances produced by the weak base learner

Algorithm – adaboost

Input: Training Data Set T, Classification Algorithm C, No of Iteration I

Read Train Data Set
Initialize weight to train data set.
For i=1 to I
Normalize the weight
Build Classification
Select Classifier with lowest error value
Update the weight
End

E. Multilayer perceptron

The most popular neural network algorithm is back-propagation algorithm. Although many types of neural networks can be used for classification purposes, the focus is on the feedforward multilayer networks or multilayer perceptrons which are the most widely studied and used neural network classifiers[8]. A multilayer perceptron (MLP) is a feed forward artificial neural network. MLP makes use of a supervised learning technique called back propagation for training. In this proposal the neural network is trained with Heart Diseases data set through the usage of feed forward neural network model and back propagation learning algorithm with momentum and variable learning rate[8]. The description of the returned propagation algorithm is specified in the above is used to educate the neural during the training process.

When comparing the accuracy, errors rate and f-measure of adaboost and multilayer perceptron. The multilayer perceptron provide higher performance than adaboost algorithm

F. Attribute Selection

In general, Attribute selection is a pre-processing step followed in machine learning. Attribute means the same as feature in weka. Attribute selection is nothing but reducing dimensionality and removing irrelevant data so that it will increase the accuracy level. In this proposal Principal component analysis evaluator is used to select the sensitive attributes. It performs principal component analysis and transformation of data. Ranker search is used as conjunction for PCA. By using Principal component analysis seven attributes has been selected as sensitive attributes to be protected from third party or attacker. Those seven attributes are sex, age, trestbps, chol, fbs, cp, num,

G. Noise generation

Noise generation is achieved through perturbation based PPDM. Additive Perturbation technique is masking the attribute values by adding noise to the original data. The noise added to the data is as large as possible from that the individual record cannot be recovered [4]. Perturbation of data is known by other names such as data noise and data distortion. Before applying perturbation, attribute selection mechanism will be carried out by using ranker search method, so that the sensitive attributes alone can be perturbated.

The additive data perturbation approach can generate the perturb data Z by adding the original data X with random noise Y. This can be represented as follows:

$$Z=X+Y$$

IV. EXPERIMENTAL RESULT

In this proposal the experiment has been accomplished using ensemble classification to classify the heart disease data and by finding the performance evaluation in the data we can get accuracy level and error rate. So these data will be suitable for privacy preserving data mining technique (PPDM). The performance evaluation of the ensemble classifier is calculated using Confusion Matrix, F-measure, precision, recall and Accuracy. In our proposal, error rate, accuracy and F-measure are used to determine the quality of classification system.

A confusion matrix contains information about actual and predicted classifications done by a classification system.

Table 4.1: Confusion Matrix

		PREDICTED CLASS	
		Positive	Negative
Actual class	Positive	True Positive (TP)	False Positive (FP)
	Negative	True Negative (TN)	False Negative (FN)

- **Accuracy (AC)** can be determined by the ratio between correctly classified instances and total number of instances.

  $$Accuracy = \frac{true\ positive + true\ negative}{tn+tp+fn+fp}$$

- **Misclassification Rate** is also known as ”Error Rate” is the ratio of the incorrectly classified instances by total number of instance or equivalent to 1 minus Accuracy.

  Error Rate= (FP+FN)/N or Error Rate =1- Accuracy

- **F-measure** is a measure of a test's accuracy. It considers both the precision and the recall.

  $$Precision = \frac{true\ positive}{true\ positive + false\ positive}$$

  $$Recall = \frac{true\ positive}{true\ positive + false\ negative}$$

  $$F\text{-}measure = \frac{2 * precision * recall}{precision + recall}$$

Table 4.2: performance of the classifiers

Evaluation criteria	**Adaboost**	**Multilayer perceptron**
Correctly classified instances	166	256
Incorrectly classified instances	137	47
Accuracy	54.79	84.49
Error Rate	0.45	0.15

By comparing the performance evaluation of adaboost and Multilayer perceptron the result analysis shows that, Multilayer perceptron gives better results than Adaboost. Hence Multilayer perceptron has been adopted. The dataset used in this experiment is heart disease dataset, which is a real dataset collected from the UCI Repository. This dataset contains 303 records and 14 attributes.

V. CONCLUSION

As the conclusion, In this system The sensitive attribute of the patients are protected by perturbation based PPDM. By making use of processes such as pre-processing followed by the ensemble classification by comparing the Adaboost and Multilayer Perceptron to classify heart disease problems as it is efficient and fast manner. Eventually noise generation through perturbation based PPDM is implemented to achieve the goal. So that the perturbed data do not corresponds to real-world record. Thus, the attacker cannot perform or make use of the sensitive records of an individual.

REFERENCES

[1] Mehmet Emre Gursoy, Ali Inan, Mehmet Ercan Nergiz, and Yucel Saygin , " Privacy-Preserving learning analytics: challenges and techniques" , IEEE transactions on learning technologies, vol. 10, no. 1, january-march 2017 , pp 68-81

[2] Jasmina D. Novakovic, Alempije Veljovic, "AdaBoost as Classifier Ensemble in Classification Problems", INFOTEH-JAHORINA Vol. 13, March 2014, pp616-620.

[3] Arshveer Kaur, "A Hybrid Approach of Privacy Preserving Data Mining using Suppression and Perturbation Techniques", International Conference on Innovative Mechanisms for Industry Applications (ICIMIA 2017), pp306-311

[4] Hamirpur, Hamirpur," International Journal of Computing, Communications and Networking" volume 4 No.1,January-march 2015

[5] R.Kalaivani ,S.Chidambaram, "additive gaussian noise based data perturbation in multi-level trust privacy preserving data mining", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.4, No.3, May 2014

[6] Hillol Kargupta,Souptik Datta1,Qi Wang, Krishnamoorthy Sivakumar , "Random-data perturbation techniques and privacy-preserving data mining" , Springer-Verlag London Ltd. 2004 Knowledge and Information Systems (2005) 7: 387–414 , pp 318-414

[7] Roman I. Batygin, Olga K. Alsova , " Software System for Different Types of Data Classification Based on the Ensemble Algorithms" , 2016 13th International Scientific-Technical Conference APEIE – 39281 , pp 506-509

[8] K. Usha Rani," Analysis of heart diseases dataset using neural network approach", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.1, No.5, September 2011