Volume 7, Issue 2, February 2017 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

# Privacy Preserving Fuzzy Based Decision Tree Classifier

**N. G. Nageswari Amma**
Assistant Professor, Department of Computer Applications, Vins Christian College of Engineering, Chunkankadai, Nagercoil, Tamil Nadu, India

**Dr. F. Ramesh Dhanaseelan**
Prof & Head, Department of Computer Applications, St.Xaviers Catholic College of Engineering, Nagercoil, Tamil Nadu, India

DOI: 10.23956/ijarcsse/V6I12/0200

***Abstract:*** ***Huge volumes of personal data are collected and sharing of these data is beneficial for machine learning and data mining applications. These data are very much important for decision making in the competitive business world. Therefore, privacy preserving has been developed to hide sensitive data. We present a privacy preserving fuzzy based decision tree classifier, which allows extracting rules from the decision tree and the sensitive data are fuzzified in order to preserve the privacy. The objective of privacy preserving data classification is to build accurate classifier without revealing private information in the data being mined. In this paper, the advantages of fuzzy logic and decision tree are used to build the privacy preserving classifier. Fuzzy logic is used to preserve the privacy in the system and decision tree is used for classification. The effectiveness of the classifier is verified by experiments on Hepatitis dataset provided by the University of California, Irvine (UCI) machine learning repository. The classification accuracy obtained using this approach is 83.63%.***

***Keywords: Privacy preserving, fuzzy logic, decision tree, hepatitis, classification***

## I. INTRODUCTION

Recently, due to the popularity of electronic data held by commercial corporations, there is an increasing demand for the privacy protection of personal information. Data mining techniques have been viewed as a threat to the sensitive content of personal information. This kind of privacy issue has led to the research for privacy preserving data mining [1]. One of the important data mining tasks is classification. The classification algorithm learns a classification model from labeled training data for the future use of classifying unseen data. There have been many privacy-preserving schemes designed for various classification algorithms, among these decision tree is the most popular method.

The knowledge representation, making decisions in case of uncertainty, and improper precision, choosing and adopting a suitable model are some of the challenges to be considered while building an intelligent system. In this paper, the advantages of fuzzy logic and decision tree are combined to develop a privacy preserving classifier. The key benefit of fuzzy logic is that it lets the designer describe the desired system behavior with simple if-then relations, thus the design time is reduced.

In most of the applications like disease prediction system, and intrusion detection system, the knowledge that describes the system performance is contained in the datasets. When the datasets contain the knowledge about the system to be designed, a decision tree promises a solution, because it can generate rules from the datasets. The decision tree is like a tree structure in which the internal nodes represent the test on the attribute and each branch represent the outcome of the test and leaf nodes hold a class label. By combining the explicit knowledge representation of fuzzy logic with the rule extraction power of decision tree, we can derive a model with higher predictive accuracy.

In many applications, privacy issues come up because their data are considered as sensitive data. As a result, normal methods for knowledge discovery process are not appropriate. Therefore, privacy preserving data mining methods are used to build models and extract patterns without revealing the private data. In this paper, fuzzy logic is used in order to preserve the privacy and decision tree is used to extract the rules and to classify the data.

## II. RELATED WORK

In the literature, many approaches have been used to build a privacy preserving classifier. Tamir, developed a system which securely mines the association rules from horizontally distributed databases. This system is very much efficient in terms of computational and communicational costs [2]. Pui and Jens developed a privacy preserving approach using ID3 decision tree learning algorithm with discrete valued attributes. They implemented algorithms such as C4.5 and C5.0 with application scope and data mining methods with mixed discrete and continuous valued attributes. They optimized the storage size and the processing time, when generating a decision tree from the samples [3]

Alka and Patel discussed a privacy preserving data mining method using the decision tree over horizontally partitioned data using untrusted third party. In their work, each and every party calculated the result and it is called as intermediate result and they send the intermediate result to untrusted third party. They also proved that the performance of privacy preserving two-layer horizontally partitioned ID3 decision tree classifier is better than the basic ID3 decision tree classifier [4].

 

Anand and Ojha used algorithms like ID3, Gain Ratio, and Gini Index for constructing a decision tree. They generated association rules to perform data generalization, data summarization and data characterization and they faced secure multiparty computation problem. They used oblivious transfer protocol for secure computation. Their work has many advantages such as trust, correctness, efficiency, and fairness [5].

Bertha, David, and Santiago discussed a privacy preserving distributed learning system based on genetic algorithms and artificial neural networks [6]. Their system solved the machine learning challenges like tackling massive databases, learning in distributed environment and preserving the privacy of sensitive data. Benny provided an overview of the new and rapidly emerging research area of cryptographic based privacy preserving data mining, the privacy preserving techniques are classified, reviewed and evaluated [7]. Lindell and Benny discussed about the various tools that can be used to solve several privacy preserving data mining problems [8]. They also proposed a secure algorithm to build a decision tree using ID3 over horizontally partitioned data between two parties using secured multi-party computation.

Classification is one of the most widespread data mining problems come across in real life. Quinlan proposed ID3 decision tree classification and showed that it is the best classification algorithm [9]. Fang and Yang [10] proposed a privacy preserving decision tree algorithm over vertically partitioned data, which is based on idea of passing control from site to site. Agrawal and Srikant [11] discussed a randomization based perturbation approach to preserve the privacy of the data. The perturbation is done individually on the data by adding some noise that is randomly drawn from a known distribution. Then the decision tree classifier is learned from the reconstructed aggregate distributions of the perturbed data.

Yaping, proposed a privacy preserving data mining approach by enabling multilevel trust among parties. They allowed more trusted data miner to access the perturbed copy of the data and disallowed malicious data miner [12]. Vaidya et al. proposed tools for privacy preserving data mining [13]. They also proposed privacy preserving decision trees over vertically partitioned datasets [14]. Du and Zhan built decision tree classifier over private data. They used only the private data to construct a decision tree classifier and they achieved better classification accuracy [15].

Keng and Ming discussed a privacy preserving Support Vector Machine (SVM) classifier [16]. In their work, the SVM trains the classifier to decide which of the training dataset support vectors are used. The classifier designed by them violates the privacy. So they post-process the classifier to preserve the privacy of the sensitive data.

Pradeep et al. proposed a fuzzy based clustering algorithm for privacy preserving data mining. They transformed the attributes to fuzzy attributes to preserve the privacy. They used ID3 and Naïve Bayes classification algorithms over three different datasets to show the effectiveness of their work [17]. Emekci et al. proposed a privacy preserving decision tree learning algorithm over multiple parties [18]. They used a decision tree algorithm in which each party does not contain any sensitive information that belongs to other party. Zhang and Zhong proposed a privacy preserving algorithm for distributed training of neural network ensembles [19]. Zahidul and Ljiljana proposed a framework for privacy preserving classification in data mining [20]

Lambodar, Narendra, and Sushruta proposed a privacy preserving classifier for partitioned data that uses genetic algorithm to obtain the classification accuracy [21]. Kato, Claude, and Soo discussed an ensemble classifier approach that finds to preserve data privacy. In their work, the resulting perturbed data is used to reduce the classification error [22].

Comparing to the works discussed above, the work discussed in this paper is different by using fuzzy based decision tree classification to construct privacy preserving classifier. The continuous variables of the dataset are fuzzified in order to preserve the privacy of sensitive data and the decision tree is constructed. The classification rules are generated from the decision tree.

## III. PROPOSED SYSTEM

The block diagram of the proposed system is illustrated in Figure 1. The major components of the system are Preprocessing, Fuzzification, Classification, Rule base, and Inference.

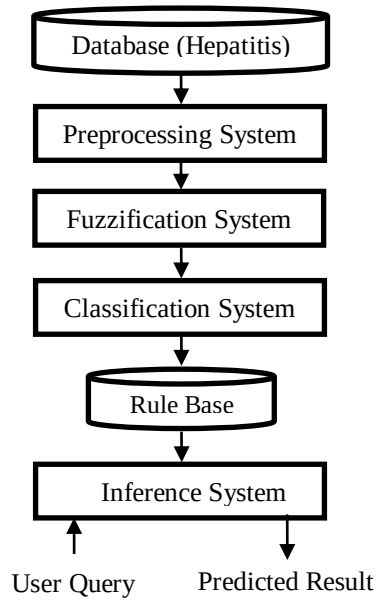


Figure 1. Proposed System Architecture

 

### 3.1 Hepatitis Database

The Hepatitis data provided by the University of California, Irvine Machine Learning Repository is used for analysis of this work. Table 1 shows the details of the Hepatitis dataset. The dataset has 19 numeric input attributes namely age, sex, steroid, antivirals, fatigue, malaise, anorexia, liver big, liver firm, spleen palpable, spiders, ascites, varices, bilirubin, alk phosphate, sgot, albumin, protime and histology. It also has the predicted attribute i.e. the class label.

Table 1. Hepatitis Dataset

| Attribute | Domain of value |
|---|---|
| Age | 10, 20, 30, 40, 50, 60, 70, 80 |
| Sex | male, female |
| Steroid | no, yes |
| Antivirals | no, yes |
| Fatigue | no, yes |
| Malaise | no, yes |
| Anorexia | no, yes |
| Liver Big | no, yes |
| Liver Firm | no, yes |
| Spleen Palpable | no, yes |
| Spiders | no, yes |
| Ascites | no, yes |
| Varices | no, yes |
| Bilirubin | 0.39, 0.80, 1.20, 2.00, 3.00, 4.00 |
| Alk Phosphate | 33, 80, 120, 160, 200, 250 |
| Sgot | 13, 100, 200, 300, 400, 500 |
| Albumin | 2.1, 3.0, 3.8, 4.5, 5.0, 6.0 |
| Protime | 10, 20, 30, 40, 50, 60, 70, 80, 90 |
| Histology | no, yes |
| Class | DIE, LIVE |

### 3.2 Preprocessing System

Preprocessing is an important step in the knowledge discovery process, as real world data tend to be incomplete, noisy, and inconsistent. Data preprocessing includes data cleaning, data integration, data transformation, and data reduction. Data cleaning routines attempt to fill in missing values, smooth out noise while identifying outliers and correct inconsistencies in the data. In this proposed work, the most probable value is used to fill in the missing values. Data transformation routines convert the data into appropriate forms for mining. Normalization is useful for classification purpose. By normalizing the input values for each attribute measured in the training tuples will speed up the learning process. In this work, the normalization technique used is min-max normalization. The min-max normalization given in equation (1) discussed in [23] is defined as follows:

$$a^1 = \frac{a - min}{max - min}(max_{new} - min_{new}) \qquad (1)$$

### 3.3 Fuzzification System

Fuzzification is the process of mapping numerical inputs into degrees to which these inputs belong to the respective fuzzy sets. To represent a fuzzy set, we express it as a function and then map the elements of the set to their degree of membership. A membership function is a mathematical function that defines a fuzzy set on the universe of discourse. The range of values of membership functions is the unit interval [0, 1]. In this work, trapezoidal membership function is used for fuzzification. Trapezoidal membership function given in equation (2) discussed in [24] consists of four points which is defined as follows:

$$f(x,a,b,c,d) = \begin{cases} 0 & when\ x < a\ and\ x > d \\ \frac{x-a}{b-a} & when\ a \le x \le b \\ 1 & when\ b \le x \le c \\ \frac{d-x}{d-c} & when\ c \le x \le d \end{cases} \qquad (2)$$

### 3.4 Classification System

The classification system is used to generate classification rules. The proposed algorithm used for classification is as follows:

Let the training samples T in the given node of decision tree and the classes are denoted as $C_1$, $C_2$,…,$C_n$.

Step1: If T contains one or more samples, all belonging to a single class $C_i$ then the decision tree for T is a leaf identifying class $C_i$.

 

Step 2: If T contains no samples, then the decision tree is again a leaf, but the class to be associated with the leaf must be determined from information other than T, such as the overall majority class in T.
Step 3: If T contains samples that belong to a mixture of classes, refine T into subsets of samples that are heading towards single-class collections of samples.
Here, T is partitioned into subsets $T_1, T_2, \ldots, T_n$ where $T_i$ contains all the samples in T that have outcome $O_i$ of the chosen test. The decision tree for T consists of a decision node identifying the test and one branch for each possible outcome.
The entropy is calculated as follows:

If S is any set of samples, let frequency ($C_i$, S) stand for the number of samples in S that belong to class $C_i$, and |S| denotes the number of samples in the set S. Then the entropy of the set S is given in equation (3):

$$Info(S)=-\sum((freq(C_i,S)/|S|)\cdot\log_2(freq(C_i,S)/|S|)) \quad (3)$$

After set T has been partitioned in accordance with n outcomes of one attribute test X, the entropy of each and every attribute is given in equation (4). The calculation for gain is given in equation (5).

$$Info_x(T) = \sum((|Ti|/|T|)\cdot Info(Ti)) \quad (4)$$

$$Gain(X) = Info(T) - Info_x(T) \quad (5)$$

The attribute with the highest gain value is selected as root node.

### 3.5 Rule Base

The classification rules generated from the decision tree are in the rule base. The knowledge in the rule base is used in decision-making. If-then rules are one of the most common forms of knowledge discovered by data mining methods. The number and the length of extracted rules tend to increase with the size of a database, making the rule set less interpretable and useful. Rule base is a knowledge base which is used to store the extracted fuzzy If-Then rules.

### 3.6 Inference System

Inference is the process of formulating the mapping from a given input to an output using fuzzy logic. The mapping then provides a basis from which decisions can be made. The process of fuzzy inference involves membership functions, fuzzy logic operators, and if-then rules. The following is the technique used as discussed in [24]:
Step 1: Fuzzify the input variables using membership functions. The membership function values are described by a continuous function.
Step 2: Apply fuzzy AND operator in the antecedent part.
Step3: Implication from antecedent to theconsequent by analyzing rules from the knowledge base. Aggregate the consequents across the rules into single output fuzzy set.
Step 4: Defuzzify the resulting set to a single number. Mean of maxima method is used for defuzzification.

## IV. EXPERIMENTAL RESULTS

The Hepatitis Dataset provided by the UCI Machine Learning Repository [25] is used for training and testing the proposed system. The distribution of dataset is given in Table 2. Among the 155 instances of data, 100 instances are used for training and 55 instances are used for testing.

Table 2. Distribution of Data

| ***Class*** | ***Live 2*** | ***Die 1*** |
|---|---|---|
| **Train** | 87 | 13 |
| **Test** | 36 | 19 |

The dataset is preprocessed and all continuous variables are fuzzified. The continuous variables in the Hepatitis dataset are age, bilirubin, alk phosphate, sgot, albumin, and protime. Trapezoid membership function is used to fuzzify the continuous variables.

The fuzzification of the continuous variables is shown in Table 3.

Table 3. Fuzzification

| **Attributes** | **Membership Functions** |
|---|---|
| Age | Young, middle-aged, old |
| Bilirubin | Up, down |
| Alk Phosphate | Low, normal, high |
| Sgot | Up, down |
| Albumin | Low, normal, high |
| Protime | Low, normal, high |

The discrete as well as the fuzzified variables are used to construct the decision tree. The constructed decision tree is shown in Figure 2. The constructed decision tree is used to generate the classification rules. The rules generated from the decision tree are shown in Figure 3. These rules are stored in the rule base. The sample rules generated are shown in Figure 3. These rules are used by the Inference system for predicting the risk of hepatitis.

 

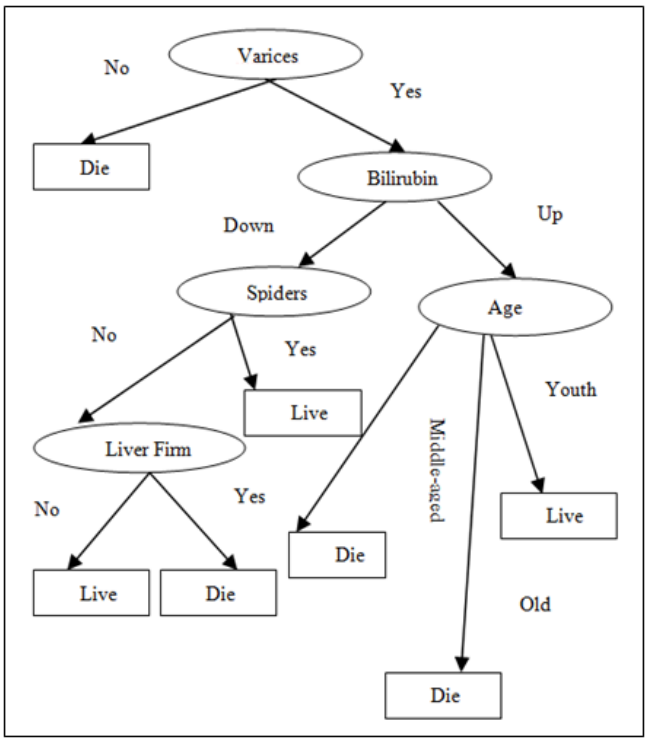


Figure 2. Constructed Decision Tree

**Rule 1: If varices is no then class is die.**
**Rule 2: If varices is yes and Bilirubin is down and spiders is no and liver firm is no then class is live.**
**Rule 3: If varices is yes and Bilirubin is down and spiders is no and liver firm is yes then class is die.**
**Rule 4: If varices is yes and Bilirubin is down and spiders is yes then class is live.**
**Rule 5: If varices is yes and Bilirubin is up and age is middle-aged then class is die.**
**Rule 6: If varices is yes and Bilirubin is up and age is old then class is die.**
**Rule 7: If varices is yes and Bilirubin is up and age is youth then class is live.**

Figure 3. Sample Rules

The performance of the system is analyzed using a confusion matrix which describes the actual and predicted value as shown in TABLE 4.

Table 4. Contingency Table

| ***Actual*** | ***Predicted*** | | |
|---|---|---|---|
| | | **Class1** | **Class2** |
| | **Class1** | TP | FN |
| | **Class2** | FP | TN |

True positives (TP) refer to the positive tuples that are correctly labeled by the classifier, while true negatives (TN) are the negative tuples that are correctly labeled by the classifier. False positives (FP) are the negative tuples that are incorrectly labeled by the classifier, while false negatives (FN) are the positive tuples that are incorrectly labeled by the classifier. The confusion matrix generated by privacy preserving fuzzy based decision tree classifier is shown in TABLE 5.

Table 5. Classification Of Testing Data

| ***Actual*** | ***Predicted*** | | |
|---|---|---|---|
| | | **Live** | **Die** |
| | **Live** | 32 | 4 |
| | **Die** | 5 | 14 |

 

The statistical measures used to test the performance of the classification are sensitivity, specificity, precision, and accuracy. Sensitivity is the statistical measure that calculates the proportion of actual positives which are correctly identified as such (e.g. the percentage of sick people who are correctly identified as having the condition). It is calculated using equation (6).

$$Sensitivity = \frac{TP}{TP + FN} \quad (6)$$

Specificity is the statistical measure that calculate the proportion of negatives which are correctly identified (e.g. the percentage of healthy people who are correctly identified as not having the condition). It is calculated using equation (7).

$$Specificity = \frac{TN}{TN + FP} \quad (7)$$

Precision is the statistical measure that calculates the degree to which further calculations show the similar results. It is calculated using equation (8).

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

Accuracy is the proportion of true results i.e. both true positives and true negatives.

Accuracy is calculated using equation (9).

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (9)$$

The performance measures are shown in TABLE 6. We compared the proposed system with the decision tree classifier and privacy preserving decision tree classifier and achieved better accuracy with the previous methods. From the performance measures, we plotted the ROC curve and it is shown in Figure 4.

Table 6. Performance Measures

| ***Measures*** | ***Decision Tree Classifier*** | ***Privacy Preserving Decision Tree Classifier*** | ***Privacy Preserving Fuzzy Based Decision Tree Classifier*** |
|---|---|---|---|
| Sensitivity | 83.33% | 86.11% | 88.88% |
| Specificity | 68.42% | 73.68% | 73.68% |
| Precision | 83.33% | 86.11% | 86.49% |
| Accuracy | 78.18% | 81.81% | 83.63% |

Figure 4 shows the receiver operating characteristic (ROC) curve for the proposed work. The ROC is a graphical plot that illustrates the performance of the classifier. The curve is drawn by plotting the true positive rate against the false positive rate at various threshold settings by using Tanagra data mining tool.

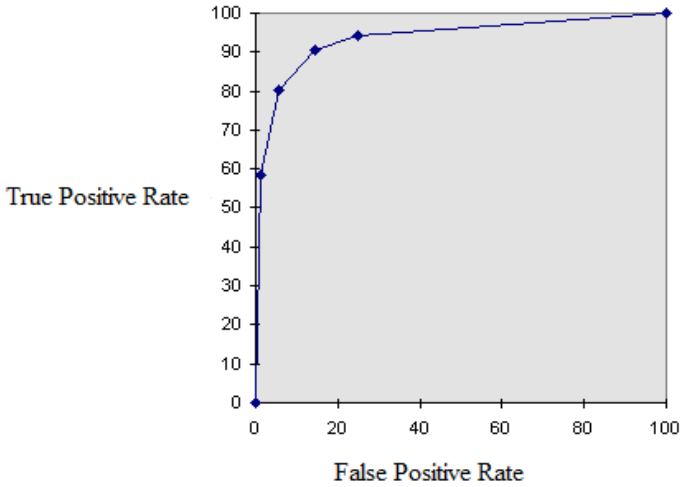


Figure 4. ROC Curve

## V. CONCLUSION

In this paper, a framework for privacy preserving fuzzy based decision tree classifier is developed for analyzing the data. The Hepatitis dataset is taken and analyzed to classify the risk of the disease. The Hepatitis data is fuzzified in order to preserve the privacy of sensitive data. The decision tree is constructed from the fuzzified data and is used to generate the rules. These rules are used for classification purpose and we obtained the classification accuracy of 83.63% using the proposed approach.

 

**REFERENCES**


[1] V. Verykios, E. Bertino(2004), State-of-the-art in Privacy preserving Data Mining, SIGMOD, vol. 33, no. 1.

[2] Tamir Tassa (2014), Secure Mining of Association Rules in Horizontally Distributed Databases", IEEE Transactions on Knowledge and Data Engineering, Volume.26, Issue.4, pp.970-983.

[3] Pui K. Fong and Jens H.Weber-Jahnke(2012), Privacy preserving Decision Tree Learning Using Unrealized Data Sets, IEEE transactions on Knowledge & Data Engineering.

[4] Alka Gangrade, Ravindra Patel(2012), Privacy Preserving Two-Layer Decision Tree Classifier for Multiparty Databases, International Journal of Computer and Information Technology (2277-0764).

[5] Anand Sharma and Vibha Ojha(2010), Implementation of Cryptography for Privacy preserving data mining, International Journal of Database Management Systems(IJDMS) Vol.2, No.3.

[6] Bertha Guijarro, David Martinez, and Santiago Fernandez(2009), Privacy Preserving Distributed Learning Based on Genetic Algorithms and Artificial Neural Networks, Proceedings of the 10th International Work Conference on Artificial Neural Networks: Part II: Distributed Computing, Artificial Intelligence, Bioinformatics, Soft Computing, and Ambient Assisted Living, PP:195-202.

[7] Benny Pinkas(2006), Cryptographic techniques for privacy-preserving data mining, ACM SIGKDD Explorations Newsletter, , vol. 4, no. 2, pp. 12-19.

[8] Yehuda Lindell, Benny Pinkas(2002), Privacy preserving data mining, Journal of Cryptology vol. 15, no. 3, pp. 177–206.

[9] J.R. Quinlan, "Induction of decision trees(1990), Jude W. Shavlik, Thomas G. Dietterich, (Eds.), Readings in Machine Learning. Morgan Kaufmann, vol. 1, pp. 81–106.

[10] Weiwei Fang, Bingru Yang(2008), Privacy Preserving Decision Tree Learning Over Vertically Partitioned Data, In Proceeding of the International Conference on Computer Science & Software Engineering.

[11] R. Agrawal and R. Srikant(2000), Privacy Preserving Data Mining, Proc. ACM SIGMOD Int'l Conf. Management of Data.

[12] Yaping Li, Minghua Chen, Qiwei Li, and Wei Zhang(2012), Enabling Multilevel Trust in Privacy Preserving Data Mining, IEEE Transactions on Knowledge and Data Engineering, Vol.24, No.9.

[13] C. Cliffton, M. Kantarcioglu, and J. Vaidya(2004), Tools for privacy preserving distributed data mining, ACM SIGKDD Explorations Newsletter, vol. 4, no. 2 pp.28-34.

[14] J. Vaidya, C. Clifton, M.Kantarcioglu, A.S Patterson(2008), Privacy-preserving decision trees over vertically partitioned data, In the Proceeding of the 19th Annual IFIP WG 11.3 Working Conference on Data and Applications Security, pp. 139–152.

[15] Wenliang Du, Zhijun Zhan(2007), Building decision tree classifier on private data, In CRPITS, , pp. 1–8.

[16] Keng Pei Lin and Ming Syan Chen(2011), On the Design and Analysis of the Privacy Preserving SVM Classifier, IEEE Transactions on Knowledge and Data Engineering, Vol.23, No.11.

[17] Pradeep Kumar, Kishore Indukuri Varma, and Ashish Sureka (2011), Fuzzy Based Clustering Algorithm for Privacy Preserving Data Mining, International Journal of Business Information Systems, Vol.7, No.1.

[18] F. Emekci , O.D. Sahin, D. Agrawal, A. El Abbadi(2007), Privacy preserving. decision tree learning over multiple parties, Data & Knowledge Engineering 63, pp. 348-361.

[19] Yuan Zhang, and Sheng Zhong(2012), A Privacy Preserving Algorithm for Distributed Training of Neural Network Ensembles, Neural Computer and Applications, Springer, DOI 10.1007/s00521-012-1000-8.

[20] Md.Zahidul Islam and Ljiljana Brankovic (2009), " A Framework for Privacy Preserving Classification in Data Mining", Artificial Intelligence Applications and Innovations III, IFIP International Federation for Information Processing, Vol.296, ISBN 978-1-4419-0220-7, Springer US.

[21] Lambodar Jena, Narendra Ku. Kamila, and Sushruta Mishra(2014), Privacy Preserving Distributed Data Mining with Evolutionary Computing", Advances in Intelligent Systems and Computing, Vol.247, pp.259-267.

[22] Kato, Claude Turner and Soo-Yeon Ji(2012), Towards a differential privacy and utility preserving machine learning classifier, Elsevier, Missouri University of science and technology, Washington DC,complex adaptive systems,publication 2.

[23] Jiawei Han, and Micheline Kamber (2006), Data Mining Concepts and Techniques, Elsevier.

[24] Rajasekaran.S, and Vijayalakshmi Pai.G.A.(2007), Neural Networks, Fuzzy Logic, and Genetic Algorithms: Synthesis and Applications, Prentice Hall of India.

[25] UCI (2009) Hepatitis Dataset, available at http://archive.ics.edu/ml/datasets.


 