ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

PRIVACY PRESERVING DATA MINING OF VERTICALLY PARTITIONED DATA IN DISTRIBUTED ENVIRONMENT- AN EXPERIMENTAL ANALYSIS

[1]DR. PREETI GULIA, [2]HEMLATA

[1]Assistant Professor,Department of Computer Science and Applications,
Maharshi Dayanand University, Rohtak, Haryana, India

[2] Research Scholar,Department of Computer Science and Applications,
Maharshi Dayanand University, Rohtak, Haryana, India

E-mail: [1] preeti.gulia81@gmail.com, [2] hemlatachahal@gmail.com

ORCID: [1] 0000-0001-8535-4016, [2] 0000-0002-6105-7399

ABSTRACT

Classification in data mining is the most pervasive problem in the distributed computing environment. It is really challenging to classify the data residing on different sites without revealing the private data to each other. However, decision tree classifier has become a good solution to reduce this problem. This paper explains how to build a privacy preserving decision tree over a vertically partitioned data. The decision tree is created on the data which is partitioned vertically and is owned by different parties along with concealing data of different parties. The proposed algorithm uses a best and most efficient splitting strategy for attributes and at the same time a semi-honest third party is also used. It helps different parties in calculation for construction of decision tree. Scalar product protocol and secure multi-party protocol are the keys behind the security and privacy of the method used. The experimental results show that the accuracy and precision of the proposed algorithm, that is suitable for distributed environment, is much higher than the algorithm which works on the centralized data and in which privacy is not needed.

Keywords: *Data Mining, Classification, Decision Tree Classifier, Privacy Preserving Data Mining, Vertically Partitioned Data*

1. INTRODUCTION

Data mining is a technique to study the data from different approaches and interpret it into meaningful information. It is sometimes referred to as knowledge discovery. To mine the raw data there are a number of techniques available. Classification is the most important technique of Data Mining. It is a machine learning process to group the data instances on the basis of some similarity index. Decision Tree Classifier (DTC) is an important classification technique to predict some value on the basis of older values. DTC assist the user to take complex decisions by breaking them into simpler and easy decisions. DTC is created on the basis of available data. The task of building decision tree becomes complex when the volume of data increase. This increase in volume, velocity and variety data leads a way to Big Data.

In this Big Data era, it is not possible to store the data at a centralized server or site. So, the data is divided and stored at many different locations. This is termed as distributed environment. The partitioning of data is division of a logical database into different independent parts. Database partitioning is generally done for increasing performance and to easily manage the database. Database can be partitioned in two ways:

- ❖ **Horizontal Partitioning**- means the data is divided horizontally i.e. each party is having all attributes but few instances.
- ❖ **Vertical Partitioning**- means the data is divided vertically i.e. each party is having all instances but few attributes.

For knowledge discovery or prediction of future values, whole data is required which is residing on different places. Due to shortage of hardware and communication channel efficiency, the computations are required to be done at the place where data resides. The distributed data should be assembled at one site and analyzed for discovering new meaning. In this scenario, security and privacy of data at one site, whose data is being transferred to another site for analysis, is the biggest challenge. In order to overcome this privacy preserving problem there should be some mechanism which preserves the data of users and at the same time uses the data for analysis. When the user's data is transferred to another site it should not be in its original form. It should be processed in such a way that it should not reveal its original value/meaning after transfer.

In this paper, both the above mentioned problems, building decision tree classifier for Big Data and users sensitive data is also preserved, are taken into consideration. The classification of data residing on different sites is studied with the concept of privacy preserving. In particular decision tree classifier is constructed on vertically partitioned data. Each party is having some attributes of the whole data along with the class label attribute. Decision tree is built with the help of semi-honest third party by following the privacy rules of the data. Secure Multi-party computation like Scalar product protocol(SPP) is used for computation. The decision tree is built using entropy and information gain.

The rest of the paper is organized as follows. Section 2 introduces the related work. In Section 3 concrete problem is formulated. Materials and methods are briefly described in section 4. In section 5 proposed work is presented in detail. It includes the proposed algorithm, procedure to implement it and results with graphical representation. Section 6 depicts the current work findings and their implications in the industry. The proposed work and its improvement over the prior work has been compared and presented in tabular formant in section 7. At the end, in section 8 the conclusions and future scope is mentioned.

2. RELATED WORK

One of the earliest works on privacy preserving data mining was done by Lindell and Pinkas [1]. They proposed the oblivious transfer protocol for horizontal partitioned data by using only ID3 algorithm. Agarwal and Srikant [2][3] proposed a solution to privacy preserving classification problem by reconstructing the whole data and used probability distributions. A new method to create a decision tree without revealing their private data was given by Wenliang Du, Zhijun Zhan [4]. Charu C. Aggarwal et.al [5] proposed a condensation approach of Privacy preserving data mining. The original data was reduced to different groups of different sizes and statistics was applied on them for analysis. The size of group was directly proportional to the level of privacy preserving.

Classification method of PPDM with secure multi-party computations was proposed by Animesh Tripathy and others [6]. They concluded that accuracy and privacy can be increased by tree pruning. Hemlata Chahal [7] gave a modified version of ID3 algorithm to create a decision tree and implemented it on a real dataset. It was useful predicting the value of the class attribute by giving other values according to the tree. A decision tree algorithm was proposed by Weiwei fang et.al [8] which was based on homomorphic encryption technology. The approach provided the privacy, security, accuracy and efficiency of data. H.R. Jhalla et.al.[9] proposed a method of PPDM based on linear transformations such as WHT. The method was applicable only on horizontal partitioned data. The experiments done on real data sets reveal that the method gave accuracy similar to K-NN classifier. Nasrin Irshad Hussain et.al [10] proposed a new method of privacy preserving of big data based on cryptographic technique in which encryption and key management was used. Rule based system was used for clustering. IPFS and KIPFS, two algorithms were presented by Huafeng Ba et.al [11]. K-annonymity and t-closeness approach was used of achieving user privacy. Many more researches have also worked in the field of privacy preserving based on association rules [12][13], clustering [14][15] and other methods [16][17][18][19].

On the basis of vast literature survey on Privacy Preserving Data Mining, it is concluded that there is a need to design a new DTC for vertically partitioned data by safeguarding the interests, sensitive and private data of the users. By considering this in mind following problem is formulated for research work.

3. PROBLEM FORMULATION

When the data is vertically partitioned in distributed environment and each user wants privacy of his raw data, it is very extract meaning from the whole data by considering it as one unit. So, in this research work, a new decision tree is built which takes the data from different sites or servers and gives decision on the consolidated data. The decision tree so developed should safeguard the raw data of users by not passing it to other users as it is. It processes the data first and then passes it to other users. A secure multi-party computation method is used to build the decision tree. The user is aware of only his own data and the consolidated result.

4. MATERIALS AND METHODS

4.1. Inosphere Dataset

The proposed privacy preserving decision tree algorithm is implemented by using the dataset taken from UCI repository [22]. The radar data of ionosphere is collected and presented in the dataset. It has 351 instances and 34 continuous attributes with one class attribute having value either “Good” or “Bad”. "Good" radar returns means that there exist some type of structure in the ionosphere, whereas "Bad" radar returns are those that do not show the evidence of any structure.

4.2. Weka 3.8

The proposed algorithm is analysed by adding the proposed algorithm to WEKA library on an Intel Core i3 M350 processor @2.27 GHz using eclipse IDE. Tool Weka is choosen as it is the most popular data mining tool used now-a-days.

4.3. Secure Multi-party Computations (SMC)

Secure Multi-party computation is a process of developing methods to mutually calculate a function for the data inputs of different parties while keeping the inputs safe and private. SMC can be understood as a part of cryptography. When a set or group of parties with private inputs wishes to compute some joint function of their inputs, it can be termed as Multi-party Computations. The computations become secure when the parties wish to preserve some security properties e.g. privacy and correctness. Security must be preserved in the face of adversarial behavior by some of the participants, or by an external party.

4.4. Best Splitting Strategy

The best splitting method used in the experiments are Entropy and Information gain. To calculate or find the best split for each node, largest information gain is calculated. The definition of entropy can be generalized for a discrete random variable X with N outcomes (not just two):

$$Entropy(X) = -\sum_{i=1}^{n} (\mathrm{p(Xi)} \log_2 \mathrm{p(Xi)})$$

Information gain is the entropy before splitting the node minus the entropy after splitting the node. It can be defined as:

$$Gain(S, A) = Entropy(S) - \sum\nolimits_{v \in Value(A)} \frac{|Sv|}{|S| Entropy(Sv)}$$

Each party can find its own information gain as each is having the class attribute. This value of information gain is either shared by the parties or given to a third semi honest party to identify the attribute with highest value as root node. Then to find the best split of the root node scalar product of the two parties can be calculated as:

$$Va.Vb = \sum_{t=1}^{N} (\mathrm{Va(i)} * \mathrm{Vb(i)})$$

This scalar product protocol enables parties to compute Va.Vb without sharing the information between each other. This process of splitting the nodes is repeated until the whole tree is created.

5. PROPOSED WORK

Secure Multiparty Computation (SMC) and Scalar Product Protocol (SPP) are used in the proposed algorithm to create decision tree. SMC is a process in which different parties with private inputs wish to compute some joint function of their inputs. The parties wish to preserve their private data which is also called privacy preserving data mining (PPDM). The privacy is preserved by semi-honest third party. The proposed PPDM algorithm [20] is presented below.

Proposed Algorithm

```
Owner1 computes Information Gain of all the attributes owned by him i.e. D1
Owner2 computes Information Gain of all the attributes owned by him i.e. D2.
A semi- honest third party initializes the attribute with highest information gain as the root of the tree.
    Create a queue Q to contain the root.
    while Q is not empty
    do {
            Pop up the first node N from Q for each attribute Ar[m] (for m=1….k)
            Evaluate splits on attribute Ar[m].
            Find the best split among Ar[m]'s.
            By using best split, split the node N into N1, N2….Ns
            for i=1…..s, do
                    add Ni to Q if Ni is not well classified
            }
```

5.1 Procedure

In this section the procedure of the experiments conducted is described in detail. The data is located over different locations or sites. For simplicity, in the experiments the data is divided in two parts vertically i.e. both the sites have same number of instances but different attributes. Both the parties want to classify their combined data but do not want to reveal the data to each other. The decision tree has to be made by taking all the attributes of both parties considering data privacy. In this regard, proposed PPDM algorithm of [20] is implemented WEKA 3.8. The procedure followed to build the decision tree is:-

- Calculation of split of each attribute and best split is chosen.
- Creation of partition using best split.

The splitting methods used in the experiment are: Entropy and Gini index. Both are defined as:

$$Entropy(X) = -\sum_{i=1}^{n} (\mathrm{p(Xi)} \log_2 \mathrm{p(Xi)})$$

$$Gain(S,A) = Entropy(S) - \sum\nolimits_{v \in Value(A)} \frac{|Sv|}{|S| Entropy(Sv)}$$

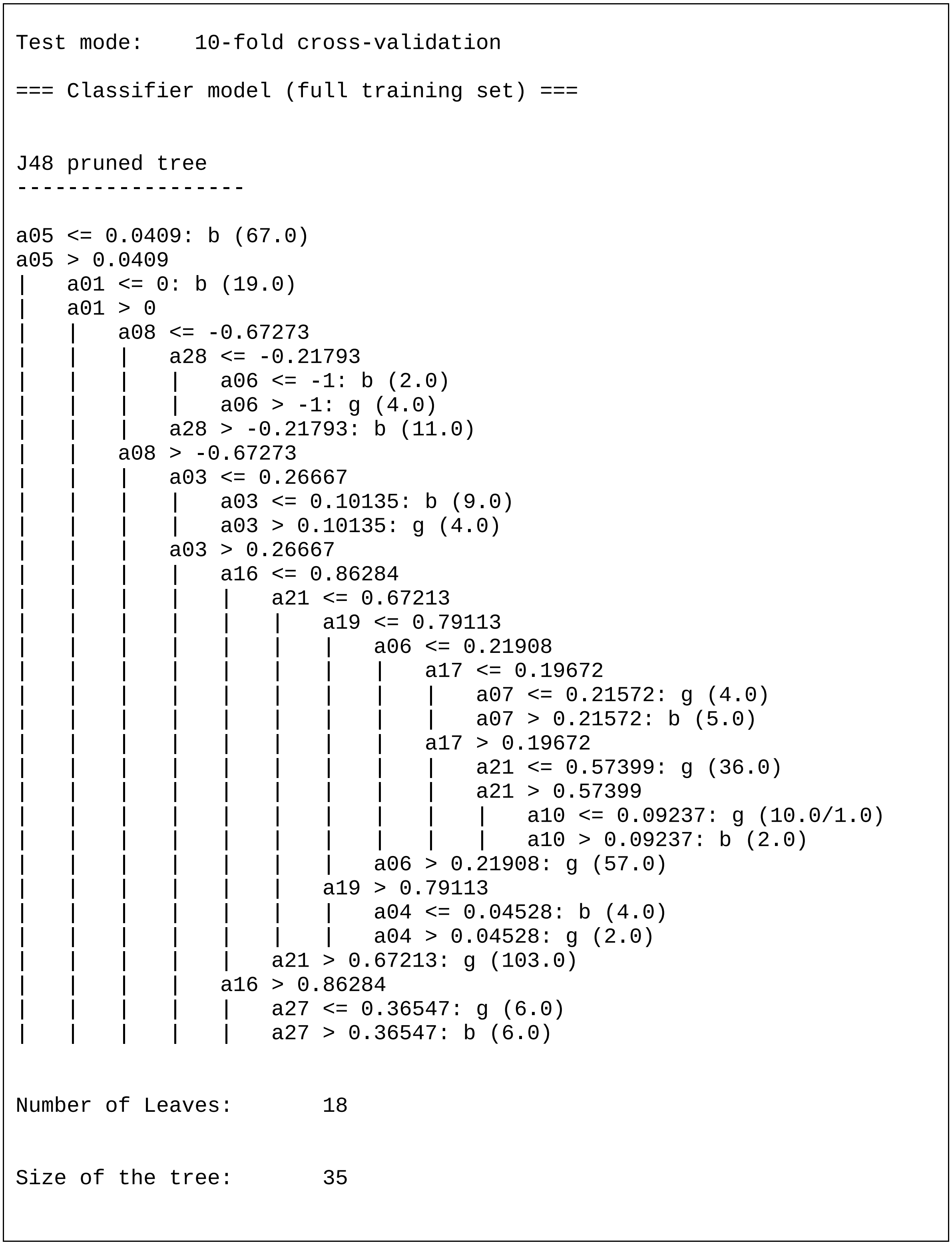
Both parties have few or selected attributes and all instances corresponding to the attributes i.e. 351 instances and 17 attributes each along with the class attribute. As both are having class attribute entropy value is calculated on its own. Each party can calculate the gain value related to its own attributes. These gain values are shared by each party and the attribute having highest gain value was chosen as root node of the tree. In this method no party can have access to other party's data only gain value is known. This procedure satisfies the privacy preserving condition of the algorithm. Likewise the tree can be generated without revealing the private data to the other party. After creating the tree it can be useful to both the parties. The algorithm follows the scalar product protocol and secure multiparty computation (SMC) protocol.

For experimental analysis of the proposed algorithm, it is compared with the existing and popular decision tree algorithm C4.5. In Weka C4.5 decision tree algorithm is implemented as J48. The rationale behind taking this existing algorithm is that it provides the accurate results as compared to other existing algorithms as per the related literature. Decision tree is created on the same dataset by using both algorithms (existing J48 and proposed PPDM). The trees are evaluated on the basis of the size and leaves of the trees. Also the accuracy and precision is used to evaluate and analyze the performance of the proposed algorithm.

5.2 Results

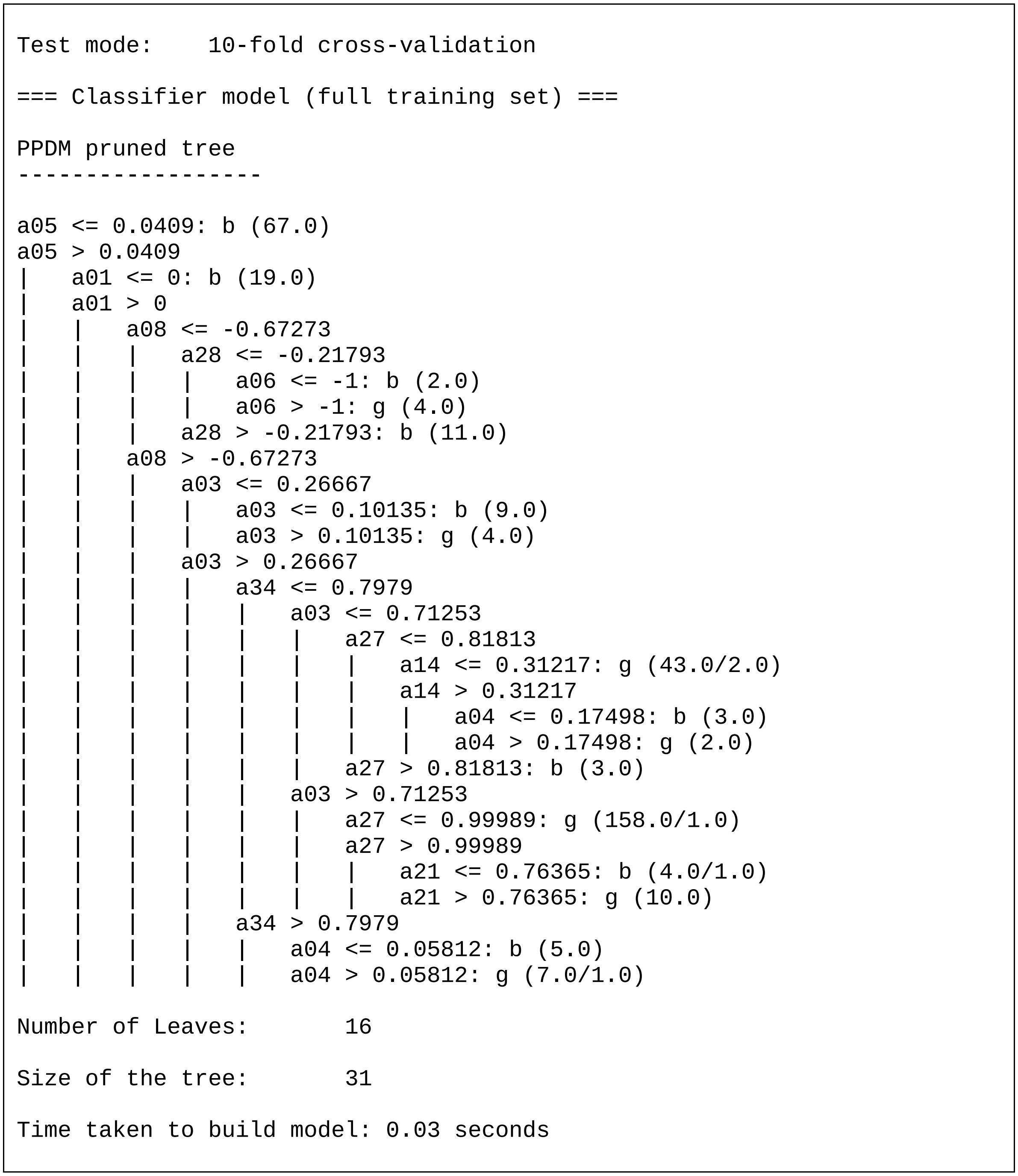
This section shows the implementation of the proposed privacy preserving algorithm [20] in Weka version 3.8. The dataset ionosphere is taken from UCI repository [22].

The experiments are conducted in a simulated distributed setting. The dataset has been randomly partitioned in order to simulate a vertical partitioning. For simplicity, in the experiments only two partitions are taken. First of all decision tree is created by using J48 algorithm(implementation of C4.5) in Weka by considering that the whole data is residing in a single site i.e. the data is not partitioned. Its results are compared with the proposed algorithm of vertically partitioned data. The algorithm uses the method of best split among the partitioned data along with information gain. Also the algorithm was so constructed to preserve the data from other party and creating tree on the combined data. The decision tree constructed in both environments (Centralized and Distributed) is shown in figure 1 and figure 2 below.



```
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===


J48 pruned tree
------------------

a05 <= 0.0409: b (67.0)
a05 > 0.0409
|   a01 <= 0: b (19.0)
|   a01 > 0
|   |   a08 <= -0.67273
|   |   |   a28 <= -0.21793
|   |   |   |   a06 <= -1: b (2.0)
|   |   |   |   a06 > -1: g (4.0)
|   |   |   a28 > -0.21793: b (11.0)
|   |   a08 > -0.67273
|   |   |   a03 <= 0.26667
|   |   |   |   a03 <= 0.10135: b (9.0)
|   |   |   |   a03 > 0.10135: g (4.0)
|   |   |   a03 > 0.26667
|   |   |   |   a16 <= 0.86284
|   |   |   |   |   a21 <= 0.67213
|   |   |   |   |   |   a19 <= 0.79113
|   |   |   |   |   |   |   a06 <= 0.21908
|   |   |   |   |   |   |   |   a17 <= 0.19672
|   |   |   |   |   |   |   |   |   a07 <= 0.21572: g (4.0)
|   |   |   |   |   |   |   |   |   a07 > 0.21572: b (5.0)
|   |   |   |   |   |   |   |   a17 > 0.19672
|   |   |   |   |   |   |   |   |   a21 <= 0.57399: g (36.0)
|   |   |   |   |   |   |   |   |   a21 > 0.57399
|   |   |   |   |   |   |   |   |   |   a10 <= 0.09237: g (10.0/1.0)
|   |   |   |   |   |   |   |   |   |   a10 > 0.09237: b (2.0)
|   |   |   |   |   |   |   a06 > 0.21908: g (57.0)
|   |   |   |   |   |   a19 > 0.79113
|   |   |   |   |   |   |   a04 <= 0.04528: b (4.0)
|   |   |   |   |   |   |   a04 > 0.04528: g (2.0)
|   |   |   |   |   a21 > 0.67213: g (103.0)
|   |   |   |   a16 > 0.86284
|   |   |   |   |   a27 <= 0.36547: g (6.0)
|   |   |   |   |   a27 > 0.36547: b (6.0)


Number of Leaves  :     18


Size of the tree :      35
```

Figure 1: Decision Tree Model of J48



```
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

PPDM pruned tree
------------------

a05 <= 0.0409: b (67.0)
a05 > 0.0409
|   a01 <= 0: b (19.0)
|   a01 > 0
|   |   a08 <= -0.67273
|   |   |   a28 <= -0.21793
|   |   |   |   a06 <= -1: b (2.0)
|   |   |   |   a06 > -1: g (4.0)
|   |   |   a28 > -0.21793: b (11.0)
|   |   a08 > -0.67273
|   |   |   a03 <= 0.26667
|   |   |   |   a03 <= 0.10135: b (9.0)
|   |   |   |   a03 > 0.10135: g (4.0)
|   |   |   a03 > 0.26667
|   |   |   |   a34 <= 0.7979
|   |   |   |   |   a03 <= 0.71253
|   |   |   |   |   |   a27 <= 0.81813
|   |   |   |   |   |   |   a14 <= 0.31217: g (43.0/2.0)
|   |   |   |   |   |   |   a14 > 0.31217
|   |   |   |   |   |   |   |   a04 <= 0.17498: b (3.0)
|   |   |   |   |   |   |   |   a04 > 0.17498: g (2.0)
|   |   |   |   |   |   a27 > 0.81813: b (3.0)
|   |   |   |   |   a03 > 0.71253
|   |   |   |   |   |   a27 <= 0.99989: g (158.0/1.0)
|   |   |   |   |   |   a27 > 0.99989
|   |   |   |   |   |   |   a21 <= 0.76365: b (4.0/1.0)
|   |   |   |   |   |   |   a21 > 0.76365: g (10.0)
|   |   |   |   a34 > 0.7979
|   |   |   |   |   a04 <= 0.05812: b (5.0)
|   |   |   |   |   a04 > 0.05812: g (7.0/1.0)

Number of Leaves:       16

Size of the tree:       31

Time taken to build model: 0.03 seconds
```

Figure 2: Decision Tree Model of proposed PPDM Algorithm

The decision tree models of both the algorithms (J48 and Proposed PPDM algorithm) show that the number of leaves in proposed algorithm is less than the leaves in J48 algorithm. Also the size of tree of proposed algorithm is less than the size of J48. It can be interpreted that the proposed algorithm is more efficient and less complex as compared to J48. The comparison of the leaves and size of tree can be shown graphically through chart 1. The model built can be visualized in the form of a tree also. Figure 3 shows the visualization of tree built by the model of J48 algorithm. Decision tree visualization of the proposed PPDM algorithm is presented in figure 4.

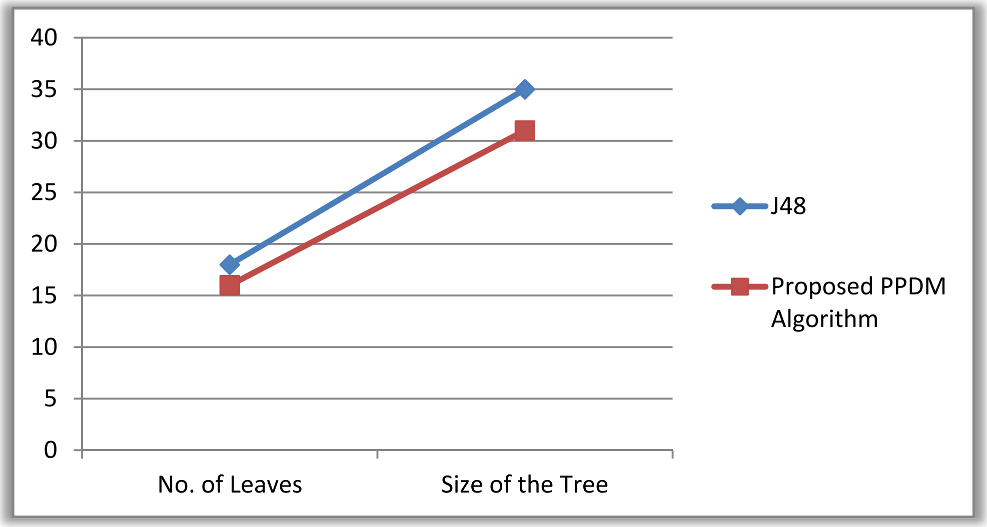


Chart 1: Comparison in terms of leaves and size of tree

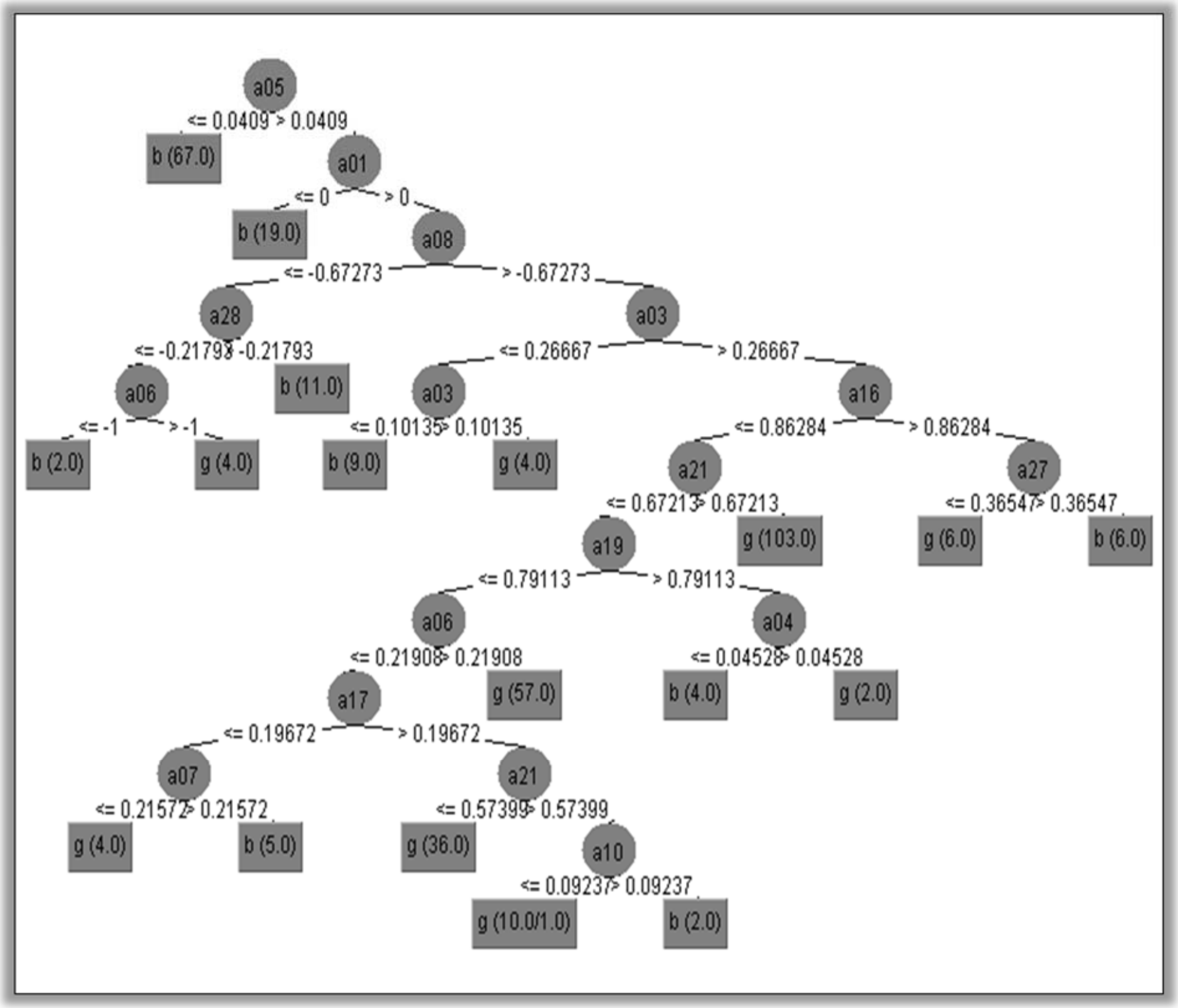


Figure 3: Visualization of Decision Tree by J48

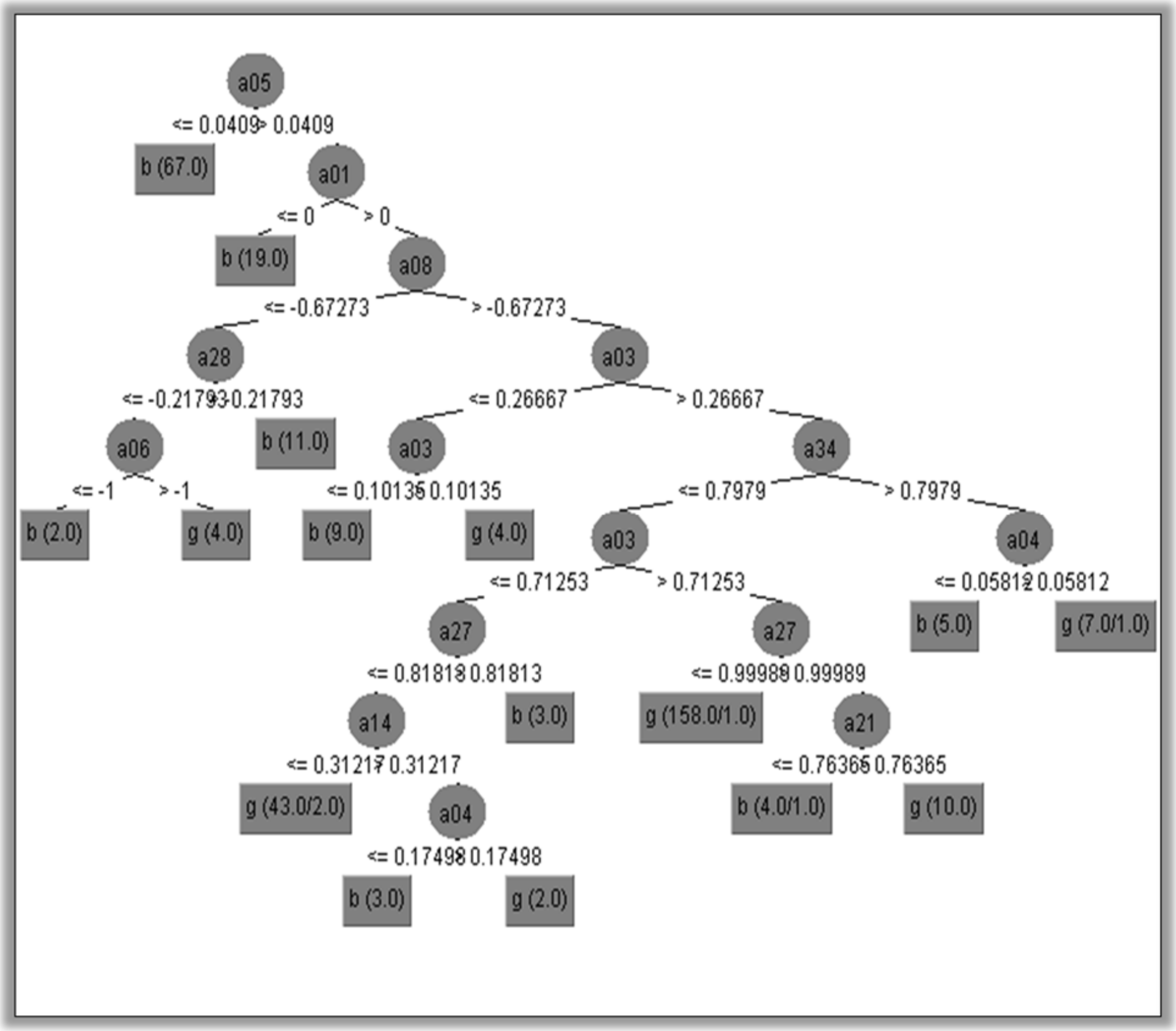


Figure 4: Visualization of Decision Tree by proposed PPDM Algorithm

Proposed algorithm [20] creates a decision tree of vertically partitioned data without disclosing the information to each participating site. Vertical partitioning is simulated on a single system, in which transmission time is not considered. The comparison parameters taken are accuracy, precision and confusion matrix. Table 1 shows the accuracy levels of both the algorithms. Accuracy of the proposed PPDM algorithm is much higher than the accuracy of built in J48 algorithm. This shows that the proposed algorithm is much accurate despite of the fact that the proposed algorithm builds the tree of vertically partitioned data and also takes care of the privacy of different users. So, in distributed scenario proposed algorithm is better than J48 in centralized environment.

Table 1: Accuracy

Algorithm	Correctly Classified Instances	Incorrectly Classified Instances	Accuracy
J48	321	30	91.453%
Proposed PPDM Algorithm	346	5	98.5755%

The classified instances by both the algorithms implemented are depicted graphically in chart 2.

From the chart it is easily observed that proposed algorithm classifies correct instances more as compared to J48. But the proposed algorithm classifies incorrect instances less than J48. With reference to both the observations it can be concluded that the proposed PPDM algorithm is better. Accuracy means how much the measured value close to the standard value. Hence, the experimental values show that the values of the proposed algorithm are closer or accurate than the standard values i.e. 100%. Accuracy as in table 1 can be graphically presented in chart 3.

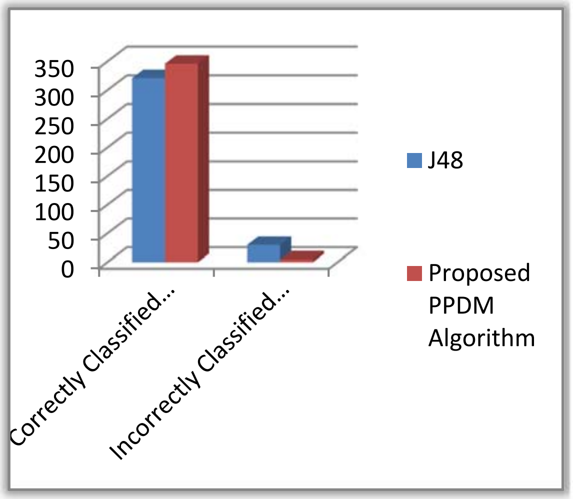


Chart 2: Classified Instances

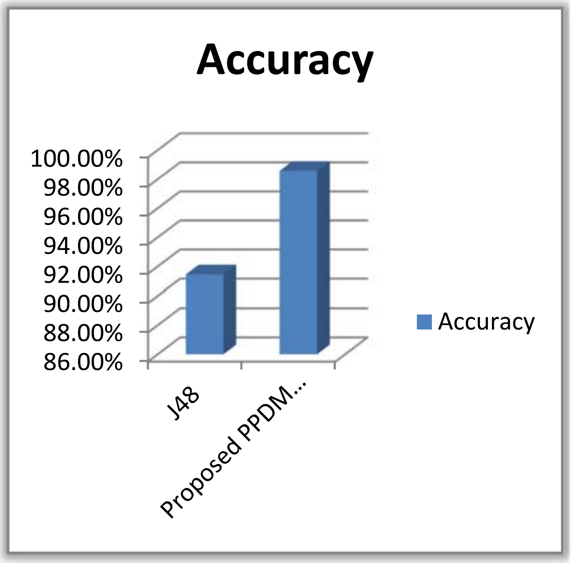


Chart 3: Accuracy

The two algorithms are also compared by taking precision as an important parameter in Table 2. J48 algorithm, an implementation of C4.5 in Weka, shows true positive rate as 0.915 and false positive rate as 0.125. The proposed algorithm gave TP Rate and FP Rate values as 0.986 and 0.022. From the values in the table it is observed that the precision of proposed algorithm is much higher than the J48 algorithm. So, it can be interpreted that the proposed algorithm is better in terms of precision. Also the proposed algorithm is very useful in privacy preserving data mining of vertically partitioned data.

Table 2: Precision

Algorithm	TP Rate	FP Rate	Precision
J48	0.915	0.125	0.915
Proposed PPDM Algorithm	0.986	0.022	0.986

The true positive rate of proposed algorithm is 0.986 and J48 is 0.915 which shows that the proposed algorithm can predict the correct values better as compared to J48. Similarly, the false positive rate of proposed PPDM algorithm is less i.e. the proposed algorithm predicts the false value less than J48. TP Rate and FP Rate of both the algorithms are shown in chart 4. The chart shows that TP Rate of proposed algorithm is higher and FP Rate of proposed algorithm is lower.

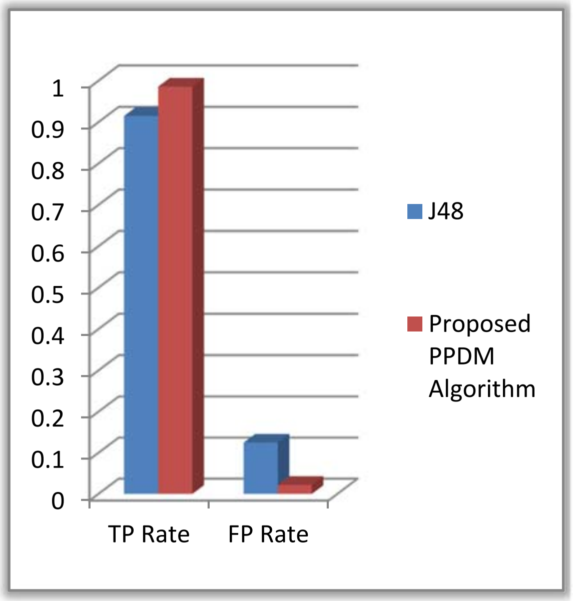


Chart 4: TP and FP Rate

Comparison of both the algorithms can be clearly shown in terms of precision. Precision means the closeness of the two measured values. The proposed algorithm excels in terms of precision also. Comparison in terms of precision is depicted in chart 5.

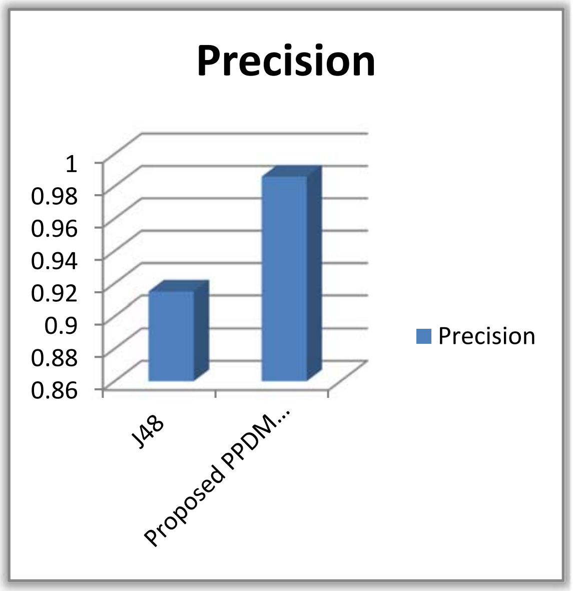


Chart 5: Precision

When the whole data is located at one central site, the confusion matrix of J48 algorithm for building decision tree classifier is shown in Table 3. J48 classified 321 instances correctly and 30 instances incorrectly. 126 instances are to be classified 'bad' but only 104 are classified as 'bad' whereas 22 instances are classified as 'good'. Also it should classify 225 instances as 'good' but classified 217 as 'good'. There are only 8 instances which should be classified as 'good' but are classified as 'bad'.

Table 3: Confusion Matrix of J48

	Bad	**Good**
Bad	104	22
Good	8	217

Similarly confusion matrix of the proposed algorithm is presented in Table 4. The proposed algorithm works in the situation when the data resides in a distributed environment i.e. at different sites. It also follows the privacy preserving protocol of all the sites. The algorithm classifies 346 instances correctly out of 351 instances. It classifies 122 instances as 'bad' but classifies 4 instances as 'good' which are to be classified as 'bad'. Also, it classifies 224 instances as 'good' which should be classified as 'good' only but classifies 1 instance as 'bad' which should be classified as 'good'. Proposed algorithm classifies 346 instances correctly and only 5 instances incorrectly which is much higher than J48 algorithm. The proposed algorithm has an edge over the earlier J48 algorithm in terms of distributed data. In the present distributed environment, proposed algorithm is much better than J48 with the benefit of privacy preserving which is of utmost demand. The proposed algorithm builds a decision tree by hiding the private data of each user with higher accuracy and precision than J48.

Table 4: Confusion Matrix of the proposed algorithm

	Bad	**Good**
Bad	122	4
Good	1	224

6. FINDINGS AND IMPLICATIONS

Findings: A novel decision tree is created which can work in distributed environment, i.e. when the data is vertically divided. The new proposed decision tree also preserves the privacy of the user's data. The proposed algorithm is validated by implementing it on a dataset and comparing the results with the pre existing J48 algorithm. The results show that the proposed algorithm has higher accuracy and precision along with privacy preserving feature.

Implications: In today's digital distributed environment, the user is unaware about the place and purpose of his data storage. So, privacy preserving is the most important technique to be implemented. The proposed algorithm is a good step in this direction. It helps the users having vertically partitioned data to take collective decisions on their data along with the security of their private data.

Most of the IT industries predict the future trend of the customers on the basis of existing data. Now-a-days huge data is available online for analysis. But due to legal implications of privacy preserving it could not be used as it is. The concerned industries can use the proposed algorithm for this purpose as it solves the privacy preserving problem. Also it takes the input data which is vertically partitioned which is the requirement of the day.

7. IMPROVEMENT FROM PRIOR WORK

Prior Wok					Current Proposed Work
S. No	**Authors**	**Publication**	**PPDM Technique used**	**Result and Accuracy**	**Improvement**
1	Charu C. Aggarwal and Philip S. Yu	EDBT, 2004	Annonymization	The proposed condensation framework led to privacy.	The proposed work leads to privacy preserving in vertically partitioned environment.
2	Yi Xia, Yirong Yang and Yun Chi	ACM, 2004	Association Rule mining	Use of non-uniform randomization factors improves the accuracy.	The proposed decision tree imroves the accuracy along eith privacy preserving.
3	Nan Zhang, Shenguan Wang and Wei Zhao	ACM, 2005	Classification	Accurate classifiers can be built and private information is not disclosed	The proposed work leads to privacy preserving in vertically partitioned environment.
4	Sheng Zhong and Zhiqlang Yang	Springer, 2007	Perturbation Technique	It gives accuracy similar to cryptographic method and is much faster than Cryptographic method.	The proposed work leads to privacy preserving in vertically partitioned environment along with higher accuracy and precision.
5	F. Emekci *, O.D. Sahin, D. Agrawal, A. El Abbadi	Elsvier, 2007	Classification method- Decision Tree	Modified algorithm was proposed to achieve privacy preserving.	New algorithm achieves privacy preserving in distributed environment.
6	Weiwei Fang and	IEEE, 2008	Classification method- Decision	It provides good capability of	The proposed work leads to privacy preserving in

	Bingru yang		Tree	preserving .privacy, accuracy and efficiency.	vertically partitioned environment.
7	Saeed Samet and Ali Miri	IEEE, 2008	Classification method- Decision Tree	The protocol was efficient and practical.	The proposed work leads to efficient and more accurate privacy preserving results.
8	Yanguang Shen, Hui Shao and Jianzhong Huang	IEEE, 2009	Privacy preserving C4.5	The method protects the privacy efficiently.	The proposed work leads to privacy preserving in vertically partitioned environment.
9	Weiwei Fang, Bingru Yang, Dingli Song, Zhigang Tang	IEEE, 2009	Classification Method - Decision Tree	Results show that the new proposed algorithm provides good capability of privacy and security.	The proposed work leads to privacy preserving in vertically partitioned environment. It shows the results more accurate and precise.
10	Sumana M, Hareesh K.S. and Shashidhara H.S.	ACM, 2010	Classification Method- Decision Tree	Semi-trusted third party commodity server was used for privacy preservation.	Semi-honest third party is used for privacy preserving in distributed environment.
11	Gopal Behera	IEEE, 2011	Decision Tree Classifier	The proposed protocol of C4.5 gave better results as compared to ID3.	The proposed algorithm gave better results than existing J48 algorithm (weka implementation of C4.5).
12	Saeed Samet and Ali Miri	Elsevier, 2012	Neural network learning method	The model is securely shared among all parties.	The results are shared among all the parties along with presrving the raw data.
13	Animesh Tripathy, Jayanti Dansana, Ranjita Mishra	ACM, 2012	Classification and Secure multi party computations	Pruning of tree improves accuracy and privacy.	Pruned Classifier is used in distributed environment.
14	Rosa Karimi Adl, Mina Askari, Ken Barker, and Reihaneh Safavi-Naini	Springer, 2012	Anonymization	Used to find the consensual privacy protection level.	The proposed work leads to privacy preserving in vertically partitioned environment.
15	Jinfei Liu, Jun Luo, Joshua Zhexue	ACM 2012	Clustering Method	Privacy of data is increased.	Privacy is inceased in vertically partitioned data.

	Huang and Li Xiong				
16	Hemlata Chahal	IJCA, 2013	Classification Method	Without revealing the bank data, the algorithm predicts the credit risk of loan seekers.	Without revealing the private data, the proposed algorithm helps in taking collective decisions.
17	Omar Abdel Wahab, Moulay Omar Hachami et-al	ACM,2014	Association Rule	Association rules queries are solved efficiently and protects t inference attacks, preserves the privacy and confidentiality.	The proposed work leads to privacy preserving in vertically partitioned environment.
18	Huafeng Ba, Xiaoming Gao, Xiaofeng Zhang and Zhenyu He	IEEE, 2014	Annonymization Method	Annonymizing the identified KIPFS, achieve better performance.	Privacy preserving is achieved in distributed environment.
19	Nasrin Irshad Hussain, Bharadwaj Choudhury and Sandip Rakshit	IJCA, 2014	Cryptographic technique	New method of privacy preserving of Big data was proposed.	Proposed work is useful in privacy preserving of Big Data in vertically partitioned data.
20	H.R.Jalla and P.N. Girija	Springer, 2016	Walsh Hadamard Transformation (WHT) and perturbation technique	The technique gave the results similar to K-Nearest Neighbour classifier	The proposed technique gave more accurate results than existing J48.

8. CONCLUSIONS AND FUTURE SCOPE

The paper presents the method of creating a decision tree classifier of vertically partitioned data with privacy preserving. The proposed algorithm creates a decision tree which does not reveal the private data of one party to another party by using secure multi-party computations. For achieving the experimental analysis objective, the proposed algorithm is implemented by taking real dataset for building the decision tree. The accuracy and precision of the tree is compared with the tree created by J48 algorithm in Weka 3.8, popular data mining software, for the centralized data residing on a single site. The experiments are conducted in simulated distributed environment on a single machine. For simplicity, only two partitions of the raw data are taken. The existing J48 algorithm is meant for the data at a single site or which is not partitioned. The proposed algorithm is meant for the vertically partitioned data which is residing in distributed environment. Both the algorithms are implemented on a real datset. The results reveal that the accuracy and precision of the proposed algorithm is much higher as compared to the built-in algorithm J48. Hence the proposed algorithm has an edge over the earlier J48 because it is used in distributed environment and conceals the private data of each party from the other party.

Other classification methods like association rules, neural networks can be used for data mining in distributed environment for future work. As a limitation, the current experiments are conducted in simulated distributed environment on a single machine. But there is a scope to conduct the experiments in real distributed environment on multiple machines so that real setup can be used. Also, further multiple partitioning of the data can be done for future work. Other cryptographic methods can be used for privacy preserving raw data.

REFRENCES:

[1] Lindell, Y. & Pinkas, B. 2000, Privacy Preserving data mining, in 'Advances in Cryptology – Crypto2000, Lecture notes in Computer Science', Vol. 1880..

[2] Agarwal, R. & Srikant, R, 2000, Privacy-preserving data mining, in 'Proceedings of the 2000 ACM SIGMOD on Management of Data', Dallas, TX USA, pp.439-450.

[3] Agarwal, D. & Aggarwal, C. 2001, On the design and quantification of privacy preserving data mining algorithms,in 'Proceedings of the 20th ACM SIGACT - SIGMOD- SIGART Symposium on Principles of Database Systems',Santa Barbara, California,USA.

[4] Wenliang Du, Zhijun Zhan, 2002, Building Decision Tree Classifier on Private Data ' Proceedings of IEEE International Conference on Data Mining', Maebashi City, Japan, Vol 14.

[5] C. Aggarwal , P.S. Yu, "A condensation approach to privacy preserving data mining", in proceedings of International Conference on Extending Database Technology (EDBT), pp. 183–199, 2004.

[6] Animesh Tripathy, Jayanti Dansana, Ranjita Mishra, "A Classification Based Framework for Privacy Preserving Data Mining", in the proceedings of ICACCI'12, August 3-5, ACM, 2012.

[7] Hemlata Chahal, "ID3 Modification and Implementation in Data Mining" International Journal of Computer Applications (0975-8887) Volume 80- No7, October 2013.

[8] Weiwei Fang, Bingru Yang, Dingli Song, Zhigang Tang, "A New Scheme on Privacy-preserving Distributed Decision-tree Mining", in the proceedings of First International Workshop on Education Technology and Computer Science, IEEE 2009.

[9] H.R.Jalla and P.N. Girija, "A Novel Approach for Horizontal Privacy Preserving Data Mining" , Advances in Intelligent Systems and Computing, pg 101-111, Springer 2016.

[10] Nasrin Irshad Hussain, Bharadwaj Choudhury and Sandip Rakshit, "A Novel Method for Preserving Privacy in Big-Data Mining", International Journal of Computer Applications(0975-8887) Volume 103- No 16, October 2014.

[11] Huafeng Ba, Xiaoming Gao, Xiaofeng Zhang and Zhenyu He, "Protecting Data Privacy from being Inferred from High Dimensional Correlated Data" in the proceedings of "IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)", 978-1-4799-4143-8/14 IEEE 2014.

[12] Omar Abdel Wahab, Moulay Omar Hachami et-al, "DARM: A Privacy-preserving Approach for Distributed Association Rules Mining on Horizontally-partitioned Data*", in the proceedings of IDEAS'14, July 07-09, ACM 2014.

[13] Vikas Ashok and Ravi Mukkamala, "Data Mining Without Data: A Novel Approach To Privacy-Preserving Collaborative Distributed Data Mining" in the proceedings of WPES'11, October 17, ACM 2011.

[14] Jinfei Liu, Jun Luo, Joshua Zhexue Huang and Li Xiong, "Privacy Preserving Distributed DBSCAN Clustering" in the proceeding of " PAIS 2012, March 30, ACM 2012.

[15] Faris Alqadah and Raj Bhatnagar, "An Effective Algorithm for Mining 3-Clusters in Vertically Partitioned Data", in the proceedings of CIKM'08, October 26-30, ACM 2008.

[16] Rosa Karimi Adl, Mina Askari, Ken Barker, and Reihaneh Safavi-Naini, "Privacy Consensus in Anonymization Systems via Game Theory", Data and Applications Security and Privacy XXVI, proceedings of 26th Annual IFIP WG 11.3 Conference, DBSec 2012, published by Springer in July 2012.

[17] Saeed Samet and Ali Miri, "Privacy Preserving ID3 using Gini Index over Horizontally Partitioned Data", 978-1-4244-1968-5/08 IEEE,2008.

[18] Sumana M, Hareesh K.S. and Shashidhara H.S., "An Approach of Private Classification

on Vertically Partitioned Data", in the proceedings of International Conference and Workshop on Emerging Trends in Technology(ICWET 2010), February 26-27, ACM 2010.

[19] Sheng Zhong and Zhiqlang Yang, "Guided perturbation: towards private and accurate mining" The VLDB Journal(2008) 17:1165-1177, Springer-Verlag 2007.

[20] Hemlata and Preeti Gulia, "Novel Algorithm for PPDM of Vertically Partitioned Data",International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 12(2017) pp. 3090-3096.

[21] J. Vaidya, C. Clifton, "Privacy Preserving Association Rule Mining in Vertically Partitioned Data", In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 639–644, 2002.

[22] https://archive.ics.uci.edu/ml/datasets/Ionosphere.