ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 9, November 2013

Privacy Preserving CART Algorithm over Vertically Partitioned Data

Raghvendra Kumar[1], Ashish Jaiswal[2], Divyarth Rai[3]

[1, 2, 3] Dept of Computer Engineering LNCT Group of College, Jabalpur, M.P., India

Abstract: Data mining classification algorithms are centralized algorithm and works on centralized database. In this information age, organizations uses distributed database. Since data mining of private data is one of the keys to success for an organization, it is a challenging task to implement data mining in distributed database. Collaboration of different organization brings mutual benefits to the party involved. So different organizations wants to collaborate and execute efficient data mining algorithm. This arises privacy issues. Organizations are unable to collaborate because the privacy of private data is not fully preserved. In this paper CART algorithm is implemented over vertically partitioned data. For efficient privacy preservation of private data, privacy preserving protocols such as scalar product, x (ln x) protocols are used.

Keywords: Privacy Preserving; CART; decision tree.

I. Introduction

Collaboration is the important key to success. It brings mutual benefit and heavy results that helps organizations in decision making. But due to the concern of privacy of data, organization hesitates to collaborate. Now privacy preservation has been an important concern since distributed data mining came into scenario. Ever since the distributed database has entered, the three problems have occurred. First problem is privacy preservation of private data. Second problem is to run centralized data mining algorithm in distributed environment. Third problem is to extract efficient result after running the centralized data mining algorithm in a secured way.

A. Data Mining

Data mining [1] has been an important aspect in the industry due to huge availability of data. Data mining is best regarded as the result of natural development of information technology. Data mining often called knowledge discovery of data as it extract important knowledge from huge amount of data. As the technology changes so is the organization. Now a day's organization uses distributed database.

B. Distributed Database

In this information age organization uses distributed database [2] because organization are geographically distributed. A distributed database is a collection of data which logically belongs to the same system but is distributed over the sites in a computer network. It is divided into three types.

i) Horizontal partitioning of database**:** Database is divided in such a way that databases D_1 and D_2 are the result of tuple sub division of a database D.

ii) ***Vertical partitioning of database:*** Vertical database are the databases that are sub divided according to the attributes. Let a database D is of size n having a_n attributes is sub divided into databases D_1 and D_2 where D_1 holding the attribute $a_1,a_2,......,a_k$ and D_2 holding the attributes a_{k+1}, $a_{k+2},......,a_n$.

Iii) Mixed partitioning of database: Here database is first horizontally partitioned then vertically partitioned or vice versa

C. Classification Rule Mining

Classification [3],[4] is a data mining or machine learning technique used to predict group membership for data instances. Classification rule mining is also called rule based classification. It uses IF-THEN RULES. The IF part of the rule is known as rule antecedent and the THEN part is known as rule consequent. These rules can be directly be mined from training set or indirectly by converting models and extracting the rule.

D. Decision Tree Classifier

It is a supervised learning method that constructs decision trees from training set data. Decision Tree Classifiers [3] are effectively used in areas like radar signal classification, character classification, medical diagnosis, experts system etc. A decision tree [1] is a flowchart like selection measures are used for the attribute that best partitions the tuples into different classes. During decision tree construction, many of the branches may reproduce noise or outliers in the training data. So tree pruning is done to improve the algorithm. It breaks the complex decision making process into a collection of simpler decisions. Decision tree algorithm has adopted greedy approach in which decision trees are constructed in top-down recursive divide-and-conquer manner. Most commonly used decision tree algorithms are Iterative Dichotomiser (ID3), C4.5 and CART. Privacy preserving decision tree algorithm has been used to solve distributed computation problem where the parties altogether build a decision tree over the data set by sharing the necessary data and preventing the exposure of private sensitive data.

E. Privacy Preserving Techniques

i) Privacy preserving has emerged data mining has emerged as a very active research area in data mining. Discovering knowledge through a combination of different databases raises security issue. Although data mining results usually do not violate privacy of individuals, it cannot be assured that an unauthorized person will not access the data is partitioned over different sites and data is not encrypted, it is impossible to derive new knowledge about the other sites. Data mining techniques try to identify regularities in data, which are unknown and hard to discover by individuals. Regularities or patterns are to be revealed over the entire data, rather than on individuals. However to find such disclosure of patters, the mining process has to access and use individual information. Techniques discussed in this paper are *Scalar Product Protocol:* The scalar product protocol[4] allows more than two parties for computation. The main goal of this protocol is to secure the private data of other parties such that a party can know its own result and data only.

ii) ***X(lnX) Protocol:*** X(lnX) protocol[5],[6] is generally used for preserving privacy for two parties. Suppose we have two parties A and B having value X_a and X_b respectively. The goal of X(lnX) protocol is to give A and B both a share of Y_a and Y_b respectively such that

$$Y_a + Y_b = (X_a + X_b)\ln(X_a + X_b)$$

II. RELATED WORKS

Classification [3], one of the machine learning techniques, has been a major breakthrough in data mining process that lead to decision tree. There are three most popular decision tree. First decision tree is ID3 [3] that uses information gain for attribute selection. Second decision tree is C4.5 [3] that uses gain ratio as its splitting attribute. Third decision tree is CART [7] that uses gini index as its splitting criteria. Privacy has been the major concern since the distributed database came into picture. Data mining classification algorithm uses centralized algorithm and are only used in centralized database. The main breakthrough was from Lindell and Pinkas [5] and Agrawal and Srikant [8]. Lindell and Pinkas have introduced secure multi party technique for classification using ID3 algorithm over horizontally partitioned dataset and showed the way in which how to classification algorithm can run in distributed database and preservation of privacy in private data has started from them. Through this path an extensive research on applying data mining classification technique in distributed database was conducted. To add more security Du and Zhan [4] has introduced a scalar protocol. Vaidya and Clifton in [9] and J. Vaidya [10] have introduced a secured way to apply ID3 classification algorithm in vertically partitioned database. Shen, Shao and yang [11] and Y. Shen, H. Shao, J. Jianzhong [12] studied C4.5 classification and introduced PPC4.5 algorithm over vertically distributed database for two parties. C4.5 algorithm has been extended to multiparty in vertically partitioned environment in [13]. In this paper we have designed privacy preserving CART over vertically distributed database.

III. DATABASE MODEL

The database is distributed over the sites are described below: There are N databases D_1, ……,D_n distributed over n number of sites S_1,……,S_n in such a way that if D_i has j attributes then all database from D_1,….. , D_{i-1} and from D_{i+1}….. D_n have j attributes.

IV. PRIVACY PRESERVING DECISION TREE BUILDING

The CART algorithm was proposed by Breiman, Friedman, Olshen and Stone in [7]. CART creates binary tree and uses Gini index as its splitting attributes. This makes CART different from other decision tree. In this paper Privacy preserving CART decision tree is generated using CART algorithm.

A. Privacy Preserving Classification Problem

Let us consider a scenario of two parties, A and B, wants to conduct classification technique. A has a private database D_a and B has a private database D_b. The two databases are union of both A and B as dataset [D_a U D_b] [4] and these two databases are joined vertically by putting D_a and D_b together so that the concentration of the j^{th} record in Da with the j^{th} record in D_b becomes the j^{th} record in [D_a U D_b]. The assumptions taken are

D_a and D_b contains same number of data records.

D_a contains some attribute for all records and D_b contains the other attributes.

Both parties share class labels of all the records and also the names of all the attributes.

B. Privacy Preserving CART Algorithm

Let N maps the current node, D = D_a U D_b maps the current database. Nattr list maps the current test attributes

The PPCART Tree(D, Nattr list) Begin

STEP 1- *Create root node R*

A computes gini index for all attributes present in D_a. Similarly B computes gini index for all attributes present in D_b. Initialize the root with minimum gini index. Attribute of minimum gini index is selected as the attribute maximizes the impurity reduction.

STEP 2-*If all records in D have same class value, and then return R as the leaf node with the specified class value*

Each record has been divided in between two parties and both parties share the class labels of all records. So if we want to determine whether the both parties remain with the same single class or not, we have to verify whether the records in D_a or D_b all belong to the same single class C or not. If they belong to the same single class, then returns the leaf node with that specific class value**.**

STEP 3-*If Nattr list is empty or the left records are less than a given value then return R as a leaf node marked with the class value assigned to the most records in S.*

Both parties share the class labels of all records and the names of all attributes, so they both know whether Nattr list is null or not. If yes, just scan dataset D_a or D_b and statistic the most frequent class, marking the leaf with the most frequent class label.

```
STEP 4- A queue Q is initialized to contain the root node.
 STEP 5- While queue Q is not empty do {
STEP 6- Pop out the first node N from Q.
STEP 7- Evaluate the gini index for each attribute.
STEP 8- Find the best split attribute.
STEP 9- If the split attribute is continual then find its partition value
STEP 10- Use the best split attribute to split node N into $N_1$, $N_2$ ... $N_n$.
STEP 11-For i=1,......,n
   {
    If all records in $N_i$ belong to the same class then return $N_i$ as a leaf node marked with its class value
Else
    add Ni to Q and go on executing the PPCART Tree (D, Nattr list)
}
 }
STEP 12-Calculate the classified mistakes of each node and carry on the tree pruning.
End
```

C. Computing the Best Split

We need to calculate the gini index for each attribute to acknowledge the best splitting attribute. Let D represent the dataset belonging to the current node N. Let C represents the Needed dataset to compute the gini index for each record in the current satisfied node. There will have two kinds of situations:

If the attributes of C and the Nattr list belong to the same dataset, either of them can individually calculate gini index. If all the attributes involved in C and the _attr list do not belong to the same party, neither party can compute the information gain ratio by itself. In this case, everyone needs to union the other party to calculate it. The following three steps will be repeated until we get information gain ratio for all attributes. Finally we choose the attribute with the greatest value as a split attribute for the current node.

i) ***Computing L(D, Nattr list)**:* L represents the logical expression that satisfies the current node N. L represents the logical expression that only involved in D_a attributes. L_b represents the logical expression that only involved in D_b's attributes.

A scan dataset D_a and produce a vector of size n. $V_a(i) = 1$ if the ith record satisfies L_a else V_a (i)=0. A may calculate the value of
vector V_a by itself. Similarly, B may also calculate the value of vector $V_{b.}$

Let V_i be a vector of size n, $V_i(n)=1$, if the nth record belongs to class i else $V_i(k)$ =0. V (i)= V_a (i) ∧ V_b (i) means the corresponding record that satisfies both L_a and L_b.

$$n$$

Scalar product $V_a \quad V_b \quad V_a(j) \; V_b(j)$ means the number of

$$j\ 1$$

record whichsatisfiesbothL_a and L_b.

$P_i \quad V_a \qquad (V_b \quad V_j)\ (V_a \quad V_j)\ V_b$ means the number of belonging to class i in partition S. Now we can compute the gini index

L(D, Nattrlist)

ii)Computing Gain:

ΔGain(D, N_attrlist)

$$G(s_{1\,j} \quad s_{2\,j} \quad \ldots\ldots.. \quad s_{nj})\ \text{L(D, Nattrlist)}$$

Where

L(D, N_attrlist) computes the gini index of each attributes $G(s_{1j}+s_{2j}+\ldots\ldots+s_{nj})$ computes the gini index of the class values.

According to scalar product protocol [4] the semi-honest third party is introduced to compute the scalar product V_a V_b without revealing privacy. The result is divided into two parts

$V_a \quad V_b \quad X_a \quad X_b$ .Two parties X and Y respectively shares

X_X and X_Y, which can guarantee that X cannot get the contents of Y and Y cannot get the contents of X, so it can preserve their privacy.

According to Xln(X) protocol [5], [6] we can obtain $\ln(X_X+X_Y)=P_X+P_Y$. X and Y respectively shares P_X and P_Y. $(X_X+X_Y)\ \ln(X_X+X_Y)=(X_X+X_Y)(P_X+P_Y)= X_XP_X+ X_YP_Y+ X_XP_Y+ X_YP_X$ where the result of X_XP_Y is divided into two parts Q_X and Q_Y respectively shared by X and Y. Similarly, the result of X_YP_X is also divided into two parts S_X and S_Y, which is also respectively shared by X and Y. A can compute $W_X = X_XP_X + Q_X$
$+ S_X$ and Y can compute $W_Y = X_YP_Y + Q_Y + S_Y$ .So $X_X+X_Y\ \ln(X_X+X_Y) = W_X+W_Y$, the result is divided into two parts W_X
and W_Y ,and respectively shared by X and Y.

Let us consider a scenario where two parties X and Y want to conduct CART algorithm. Both parties share the class value (play).

TABLE I
PRIVATE DATA OF PARTY X

SL NO.	OUTLOOK	TEMPARATURE	PLAY
1	SUNNY	85	NO
2	SUNNY	80	NO
3	OVERCAST	83	YES
4	RAINY	70	YES
5	RAINY	68	YES
6	RAINY	65	NO
7	OVERCAST	64	YES

8	SUNNY	72	NO
9	SUNNY	69	YES
10	RAINY	75	YES
11	SUNNY	75	YES
12	OVERCAST	72	YES
13	OVERCAST	81	YES
14	RAINY	71	NO

TABLE II
PRIVATE DATA OF PARTY Y

SL NO	HUMIDITY	WINDY	PLAY
1	85	FALSE	NO
2	90	TRUE	NO
3	86	FALSE	YES
4	96	FALSE	YES
5	80	FALSE	YES
6	70	TRUE	NO
7	65	TRUE	YES
8	95	FALSE	NO
9	70	FALSE	YES
10	80	FALSE	YES
11	70	TRUE	YES
12	90	TRUE	YES
13	75	FALSE	YES
14	91	TRUE	NO

Now both parties have their own private data and wishes the other party should not know their private data. So the problems occurred is to find the gini index and differential gini without the knowledge of second party. Let R be the requirement and it is divided into two subsets R_X and R_Y where R_X is the subset of requirement involving party X attributes and R_Y is the subset of requirement involving party

Y attributes. Let us consider two vector V_X and V_Y are of size n respectively. $V_X(t)=1$ and $V_Y(t)=1$ if t^th^ record satisfies R_X and R_Y respectively. else $V_X(t)=0$ and $V_Y(t)=0$. Let us consider another vector V_B to know if t attribute belong to a class C or not. If $V_B(C)=1$ then attributes being to class C else $V_B(C)=0$. V is a non zero entry where $V(t)=V_X(t) \wedge V_Y(t)$ $(t=1,2,\ldots.,n)$ means V(t) is satisfying both R_X and R_Y. Now party X and Y can compute their own private data by the following formulas

For computing scalar product of V_X and V_Y

$$V_X \quad V_Y \qquad \overset{n}{\underset{t\ 1}{}} V_X(t) * V_Y(t)$$

For computing P_C which means calculating number of occurrences of class C in a partition p is

$$P_c \quad V_X \quad (V_Y \quad V_C) \ (V_X \quad V_C) \ V_Y$$

For the computation of gini index party X has a vector V_X and party Y has a vector V_Y both of size n. First the attribute "outlook" is computed. In outlook attribute there are three attributes "sunny", "overcast", "rainy". As we know gini index makes binary splits so he have to combine the three attributes and the combination of attributes having minimum gini index is chosen as splitting attributes. So

$V_{X(outlook)}$(sunny,overcast)=$(1,1,1,0,0,0,1,1.1.0,1,1,1,0)^T$ $V_{X(outlook)}$ ((sunny,overcast)-no)=$(1,1,0,0,0,0,0,1,0,0,0,0,0,0)^T$ $V_{X(outlook)}$ ((sunny,overcast)-yes)=$(0,0,1,0,0,0,1,0,1,0,1,1,1,0)^T$ $Vx_{(outlook)}$ (sunny,rainy)=$(1,1,0,1,1,1,0,1,1,1,1,0,0,1)^T$

$V_{X(\text{outlook})}$ ((sunny,rainy)-no)=$(1,1,0,0,0,1,0,1,0,0,0,0,0,1)^T$ $V_{X(\text{outlook})}$ ((sunny,rainy)-yes)=$(0,0,0,1,1,0,0,0,1,1,1,0,0,0)^T$
$V_{X(\text{outlook})}$ (overcast,rainy)=$(0,0,1,1,1,1,1,0,0,1,0,1,1,1)^T$ $V_{X(\text{outlook})}$ ((overcast,rainy)-no)=$(0,0,0,0,0,1,0,0,0,0,0,0,0,1)^T$ $V_{X(\text{outlook})}$ ((overcast,rainy)-yes)=$(0,0,1,1,1,0,1,0,0,1,0,1,1,0)^T$ Temparature attributes is divided by the group ">75" and "<=75" since gini index doest work in real numbers. So $V_{X(\text{temparature})}$(>75)=$(1,1,1,0,0,0,0,0,0,0,0,0,1,0)^T$
$V_{X(\text{temparature})}$((>75)-no)=$(1,1,0,0,0,0,0,0,0,0,0,0,0,0)^T$ $V_{X(\text{temparature})}$((>75)-yes)=$(0,0,1,0,0,0,0,0,0,0,0,0,1,0)^T$
$V_{X(\text{temparature})}$(<=75)=$(0,0,0,1,1,1,1,1,1,1,1,0,0,1)^T$
$V_{X(\text{temparature})}$((<=75)-no)=$(0,0,0,0,0,1,0,1,0,0,0,0,0,1)^T$ $V_{X(\text{temparature})}$((<=75)-yes)=$(0,0,0,1,1,0,1,0,1,1,1,1,0,0)^T$

Again humidity attribute contains real number data so it is also divided into ">75" and "<=75".
$V_{Y(\text{humidity})}$(>75)=$(1,1,1,1,1,0,0,1,0,1,0,1,0,1)^T$ $V_{Y(\text{humidity})}$((>75)-no)=$(1,1,0,0,0,0,0,1,0,0,0,0,0,1)^T$ $V_{Y(\text{humidity})}$((>75)-yes)=$(0,0,1,1,1,0,0,0,0,1,0,1,0,0)^T$ $V_{Y(\text{humidity})}$(<=75)=$(0,0,0,0,0,1,1,0,1,0,1,0,1,0)^T$
$V_{Y(\text{humidity})}$((<=75)-no)=$(0,0,0,0,0,1,0,0,0,0,0,0,0,0)^T$ $V_{Y(\text{humidity})}$((<=75)-yes)=$(0,0,0,0,0,0,1,0,1,0,1,0,1,0)^T$
For windy attribute
$V_{Y(\text{windy})}$(false)=$(1,0,1,1,1,0,0,1,1,1,0,0,1,0)^T$ $V_{Y(\text{windy})}$((false)-no)=$(1,0,0,0,0,0,0,1,0,0,0,0,0,0)^T$
$V_{Y(\text{windy})}$((false)-yes)=$(0,0,1,1,1,0,0,0,1,1,0,0,1,0)^T$ $V_{Y(\text{windy})}$(true)=$(0,1,0,0,0,1,1,0,0,0,1,1,0,1)^T$
$V_{Y(\text{windy})}$((true)-no)=$(0,1,0,0,0,1,0,0,0,0,1,1,0,0)^T$ $V_{Y(\text{windy})}$((true)-yes)=$(0,0,0,0,0,0,1,0,0,0,1,1,0,0)^T$ For class attribute $V_{(\text{play})}$(yes)=$(0,0,1,1,1,0,1,0,1,1,1,1,1,0)^T$ $V_{(\text{play})}$(no)=$(1,1,0,0,0,1,0,1,0,0,0,0,0,1)^T$
The attribute having minimum gini index is selected as splitting point. So outlook (rainy, sunny) is selected as splitting point as $\text{GiniIndex}_{\text{outlook}}$(rainy, sunny) is the minimum among others. Party X and party Y exchange the information and it is concluded that attribute outlook with group rainy and sunny is selected as root node since it has least gini index. Suppose let outlook be rainy and windy is false then party X will compute vector
$V_{X(\text{outlook})}$(rainy)=$(0,0,0,1,1,1,0,0,0,1,0,0,0,1)^T$
$V_{X(\text{outlook})}$((rainy)-no)=$(0,0,0,0,0,1,0,0,0,0,0,0,0,1)^T$
$V_{X(\text{outlook})}$((rainy)-yes)=$(0,0,0,1,1,0,0,0,0,1,0,0,0,0)^T$ Party Y will compute vector
$V_{Y(\text{windy})}$(false)=$(1,0,1,1,1,0,0,1,1,1,0,0,1,0)^T$
Both parties have to use scalar product protocol to calculate outlook=rainy and windy= false. But to add more security xlnx protocol is also used such that

$$(X_X+X_Y)\ \ln(X_X+X_Y)=(X_X+X_Y)(P_X+P_Y)= X_XP_X+ X_YP_Y+ X_XP_Y+ X_YP_X$$

where the result of X_XP_Y is divided into two parts Q_X and Q_Y respectively shared by X and Y.
Similarly, the result of X_YP_X is also divided into two parts S_X and S_Y, which is also respectively shared by X and Y.
A can compute $W_X = X_XP_X + Q_X + S_X$ and Y can compute $W_Y = X_YP_Y + Q_Y + S_Y$ .So $X_X+X_Y \ln(X_X+X_Y) = W_X+W_Y$, the result is divided into two parts W_X and W_Y ,and respectively shared by X and Y

Correctly Classified Instances	524	89.1156%
Incorrectly Classified Instances	34	5.7823%
Kappa Statistic	0.8549	
Mean Absolute Error	0.0305	
Root Mean Squared Error	0.1745	
Root Relative Squared Error	55.3612%	
Unclassified Instances	30	5.102%
Total Number of Instances	588	

Fig. 2. ID3 Evaluation on test split

TABLE III
ID3 DETAILED ACCURACY BY CLASS

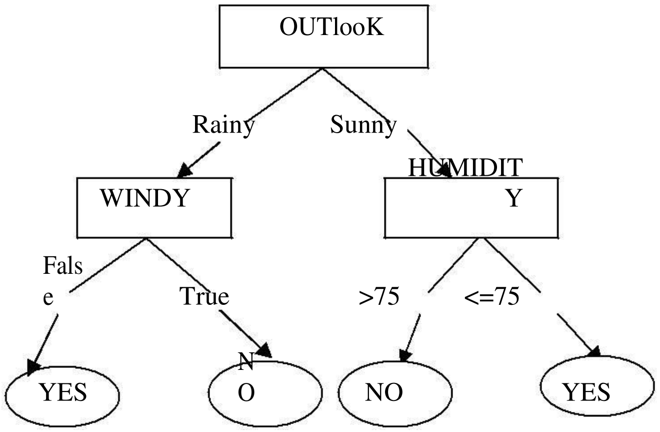

Fig. 1. CART decision tree

V. EXPERIMENTATION AND RESULTS

CART uses gini index and construct binary tree that closely model more balanced splits. Moreover it is superior to ID3 and C4.5. To prove that Weka software tool is used. Weka is an open source data mining tool. A database of car evaluation was taken for comparison purpose among the three popular trees. The results of ID3 are shown below

Correctly Classified Instances	524	89.1156%
Incorrectly Classified Instances	34	5.7823%
Kappa Statistic	0.8549	
Mean Absolute Error	0.0305	
Root Mean Squared Error	0.1745	
Root Relative Squared Error	55.3612%	
Unclassified Instances	30	5.102%
Total Number of Instances	588	

Fig. 2. ID3 Evaluation on test split

TABLE III
ID3 DETAILED ACCURACY BY CLASS

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.968	0.087	0.968	0.968	0.968	0.947	Unacc
	0.85	0.025	0.903	0.85	0.876	0.86	Acc.
	0.938	0.013	0.682	0.938	0.789	0.82	Good
	0.857	0.006	0.8	0.857	0.828	0.813	Vgood

Weighted Avg.	0.939	0.069	0.942	0.939	0.94	0.921	

a	b	c	d	⅂Γ classified as
395	11	2	0	a = unacc
13	102	3	2	b = acc
0	0	15	1	c = good
0	0	2	12	d= vgood

Fig.3. ID3 confusion matrix

The results of C4.5 are shown below

Correctly Classified Instances	535	90.9864%
Incorrectly Classified Instances	53	9.0136%
Kappa Statistic	0.8088	
Mean Absolute Error	0.0509	
Root Mean Squared Error	0.1883	
Root Relative Squared Error	55.5396%	
Total Number of Instances	588	

.

Fig. 4. C4.5 Evaluation on test split

C4.5 DETAILED ACCURACY BY CLASS

	TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**ROC Area**	**Class**
	0.946	0.045	0.98	0.946	0.963	0.972	Unacc
	0.883	0.06	0.818	0.883	0.849	0.95	Acc
	0.522	0.014	0.6	0.522	0.558	0.964	Good
	0.789	0.018	0.6	0.789	0.682	0.94	Vgood
Weighted Avg.	0.91	0.046	0.915	0.91	0.911	0.965	

a	b	c	d	– –classified as
387	20	2	0	a = unacc
8	121	3	5	b = acc
0	6	12	5	c = good
0	1	3	15	d = vgood

Fig.5. C4.5 confusion matrix

The results of CART are shown below

Correctly Classified Instances	571	97.1088%
Incorrectly Classified Instances	17	2.8912%
Kappa Statistic	0.9378	
Mean Absolute Error	0.0175	
Root Mean Squared Error	0.1109	
Relative Absolute Error	7.6337%	
Root Relative Squared Error	32.7262%	
Total Number of Instances	588	

Fig. 6. CART Evaluation on test split

Table V CART DETAILED ACCURACY BY CLASS

	TP Rate	FP Rate	Precisio N	Recal l	F-Meas ure	ROC Area	Class
	0.983	0.011	0.995	0.983	0.989	0.991	Unacc
	0.956	0.016	0.949	0.956	0.953	0.989	Acc
	0.826	0.007	0.826	0.826	0.826	0.996	Good
	1	0.007	0.826	1	0.905	0.998	Vgood
Weighted Avg.	0.971	0.012	0.972	0.971	0.971	0.991	

VI. CONCLUSION AND FUTURE WORK

This paper has showed the way for two parties to collaborate and use of CART decision tree algorithm which is better than other popular decision tree and closely model for balanced splits that brings heavy result to the party involved. Privacy of private data is fully preserved as privacy preserving protocol, that is, scalar product protocol is used. To add more security for preserving the privacy of private data, XlnX protocol is also used. So data leakage is zero.This tree can be used by different organizaion for accurate results.

REFERENCES

[1] J. Han and M. Kamber, Data Mining: Concepts and Techniques, 2nd ed., Morgan Kaufmann , New York, Elsevier, 2009.
[2] S. Ceri and G. Pelagatti, Distributed Databases: Principles and Systems, Interntional Edition, Singapore, 1984.
[3] J. R. Quinlan, "Induction of decision trees," Machine Learning, vol. 1, pp. 81–106, 1986.
[4] W. Du and Z. Zhan, "Building Decision Tree Classifier on Private Data," in proc. IEEE International Conference on Privacy, Security and Data Mining, pp. 1-8. IEEE Press, Darlinghurst, 2002.
[5] Y. Lindell, B. Pinkas, "Privacy preserving data mining", Advances in Cryptology-CRYPTO 2000, Lecture Notes in Computer Science Vol. 1880, pp 36-54, Aug. 2000.
[6] B. Pinkas, "Cryptographic Techniques for Privacy Preserving Data Mining," SIGKDD Explorations, vol. 4, pp. 12-19, 2002.
[7] L. Breiman, J. Friedman, R. Olshen and C. Stone, Classification and Regression Trees, Wadsworth International Group, 1984.
[8]R. Agrawal, and R. Srikant, "Privacy-preserving data mining," in proc. 2000 ACM SIGMOD on Management of Data, 2000, pp. 439–450.
[9]J. Vaidya, and C. Clifton, "Privacy preserving association rule mining in vertically partitioned data," in proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002, pp 639-644.
[10]J. Vaidya, "Privacy Preserving Data Mining over Vertically Partitioned Data", PhD thesis, Purdue University, 2004

[11]Y. Shen, H. Shao, L. Yang, “Privacy Preserving C4.5 Algorithm over Vertically Distributed Datasets,” in proc. 2009 International Conference on Networks Security, Wireless Communications and Trusted Computing, 2009, pp 446-448.

[12]Y. Shen, H. Shao, J. Jianzhong, “Research on Privacy Preserving Distributed C4.5 Algorithm”, in proc. 2009 Third International Symposium on Intelligent Information Technology Application Workshops, 2009, pp 216-218.

[13] A. Gangrade, R. Patel, “Building Privacy-Preserving C4.5 Decision Tree Classifier on Multiparties”, International Journal on Computer Science and Engineering, Vol.1, pp 199-205,2009.