Privacy Preserving Decision Tree Classification on Horizontal Partition Data

Kamini D. Tandel
Shri S'ad Vidya Mandal Institute of Technology
Bharuch, Gujarat, India

Jignasa N. Patel
Shri S'ad Vidya Mandal Institute of Technology
Bharuch, Gujarat, India

Abstract: **In distributed environment maintaining individual data privacy is a major issue. To generate global decision classifier among multiple parties, need to share some information with each other that arise privacy issue. Privacy preserving decision tree algorithm solve such privacy issue in distributed environment that build a global classification tree among horizontally distributed dataset and not disclosed their private sensitive data among different parties. This paper proposed a solution for privacy preserving C4.5 algorithm that deal with both discrete and continuous attribute values and use Advanced Encryption Standard Protocol to preserve privacy. Encryption/ decryption speed of algorithm is also matter. To reduce the encryption/decryption speed modified AES is used.**

Keywords— Data mining, Decision tree, Distributed Database, Privacy Preservation.

I. INTRODUCTION

Data mining means extract hidden knowledge from large amount of database and when we are dealing with mining process, privacy is the major issue. Privacy preservation in data mining protects sensitive information in database and also maintains data utility. We need to maintain the privacy of the data without sacrificing the utility of the data. Whatever data we are going to protect that generate the same result as the normal data. So the main goal of the privacy preservation in data mining is the reduce the risk of misuse of data and at the same time produce the same results as that produced in the absence of such privacy preserving techniques [1]. We can apply privacy preservation techniques at different stages of data mining process. From data collection process to generation of knowledge stage we can apply privacy techniques.

Classification is the most important task in data mining, which predict the class label of previously unknown instance. Here, we focus on decision tree classifier that is one of most popular classification technique. That follows the supervised learning approach and build classification tree based on training dataset and after that test attributes are introduced for classification of test data [2]. Decision tree is tree type structure and its leaf node represent the result of classification or class label, and its non leaf nodes represent the normal testing attributes. Different decision trees ID3, C4.5, CART uses different splitting attribute selection measure like information gain, gain ratio, and gini index respectively. ID3 only deal with the discrete value and cannot handle missing value and pruning is not performed, while C4.5 can handle discrete and continuous value and cost based pruning is done. CART have all features like C4.5 and error based pruning is performed but CART may become unstable if any changes occurred in training data and it splits only by one variable. Random Decision Tree classifier randomly chooses the variable to construct the tree that does not predict which attribute values most predictive of the class label. Among all these our propose system used the C4.5 decision tree classifier.

Privacy preservation techniques we applied based on input data source. If data stored in one single location/machine then it is a centralized database environment. And if data distributed among different parties/machines then its distributed environment. In distributed database data fragmentation is applied and store column wise fragmented data in vertical partition database and raw wise fragmented data in horizontal partition database. If we are dealing distributed environment jointly data mining tasks are performed and SMC is most suitable at this stage. To construct the global decision tree classifier between multiple parties we need to share some information that time individual data privacy needs to preserve. Existing work use the SMC techniques to easily take care about the privacy between different parties. Here we are using AES algorithm to deal with privacy preservation. Encryption and decryption speed is also major factor of privacy preservation algorithm. AES algorithm takes more time to encrypt and decrypt the data so we are using modified AES algorithm that will reduce the encryption and decryption time.

II. RELATED WORK

In distributed environment privacy violation occurred when new instance need to classify. Privacy preservation via anonymization technique suffers from homogeneity attack, background attack and heavy information loss. In perturbation technique original data reconstruction is very difficult. Randomization technique treats all data with equal priority and in condensation technique information loss occurred due to condensation of large amount of data. Cryptography technique provides good security with extra communication cost for encryption and decryption. And in distributed database cryptography techniques are most suitable. In [3] Lindell and Pinkas described the first work done on ID3 decision tree on horizontal partition data. It worked on two parties only and use the Yao's two party protocol and oblivious polynomial equation protocol for provide privacy. In [4] author developed the privacy preserving ID3 algorithm for multiple parties over horizontal partition. It used the secure multiparty computation sub protocol and homomorphic encryption scheme for privacy. Privacy preservation on vertical partition data on a decision tree work on multiple parties that again used the SMC protocol [5]. In [6] gini index is used to find the best information gain for ID3 on horizontal partition data. CART algorithm developed in [2], which used the SMC protocol to provide privacy. Data perturbation technique is used to provide privacy in [7], which used the random noise addition scheme to protect the data. In [8] author also used the perturbation and randomization based techniques to find the partially corrupted data. Advanced Encryption standard (AES) is used to provide privacy, which generate decision tree from encrypted data and generate accurate classification result [9]. Data complementation approach used to provide privacy and decision tree is generated from transformed data [10]. Random decision tree framework developed in [11], which used the homomorphic encryption scheme for provide privacy and work on both partitions horizontal and vertical that reduce information leakage. RDT framework chooses the random variable to build the decision tree but it has some privacy violation. RDT framework shares the schema structure among different parties so that other party can easily guess the information if number of new instances is classified by same leaf node. And that become easy for other party that figure out the branch structure of the tree. RDT framework generates accurate models with much smaller cost and accuracy increase as number of tree construction also increases. And that become the drawback of this approach, if the number of tree increases the complexity of structure and time complexity is also increase. The proposed C4.5 algorithm generate single decision tree among different party so it reduce the time and before the completion of tree error pruning is done so it increase the classification accuracy.

III. PRIVACY PRESERVATION TECHNIQUES

Privacy preserving approach considers the specific parameter like cost, complexity, utility, performance, security, etc. It's very difficult to design an algorithm that follows the all specific parameter. This paper follows the security enhancement parameter at any cost using AES. All these existing work used the SMC (Secure Multiparty Computation) protocol on distributed environment. Where more than two parties are involved, we are using SMC protocol. The basic idea behind the SMC is that after the secure computation of the process, each party knows its own input and result only. SMC works on two adversarial models: (1) Semi-Honest model (2) Malicious model. Semi-honest model that follow the protocol specification and try to get additional information from analyzing the messages received during the protocol execution. In Malicious model does not follow the protocol specification. It is easier to design model for semi-honest adversaries than malicious adversaries. Homomorphic encryption can be done on already encrypted data and same result can be obtained as an original data. Cryptography techniques provide secure result and less privacy leakage, but increase the encryption and decryption overhead. And it becomes less efficient for larger dataset and involved more number of parties because of extra communication overhead. Cryptography techniques divided into two parts: asymmetric key (public key) algorithm and symmetric key (private key) algorithm and block cipher and stream cipher are the part of symmetric key. In asymmetric key encryption two keys are used public key for encryption and private key for decryption. Block cipher is part of symmetric key encryption that encrypts the data in block rather than a bit at a time. If we are more concern about the privacy we can use block cipher that provide more security but with some additional cost. The proposed approach used AES algorithm to encrypt the data that first encrypt the data at different site and that encrypted data send to trusted third party to generate the global decision tree classifier.

AES block cipher algorithm has 128 bits fixed block length. The AES algorithm is consist of four main functions that executing nth times depends upon the key length. It has three different key lengths 128, 192 and 256 bits. It has 10 repetition rounds processing for 128 bit key, 12 rounds for 192 bit key and 14 rounds for 256 bit key. AES is operating on (4*4) state array and encryption decryption process done on state array. Initially state array consists of input data and it will keep changing in each round until final encrypted data is generated. And in decryption process reverse operation is done. All repeated rounds perform the same except the last round and in each round four different stages are there: add round key, sub byte operation, shift row transformation and mix column operation.

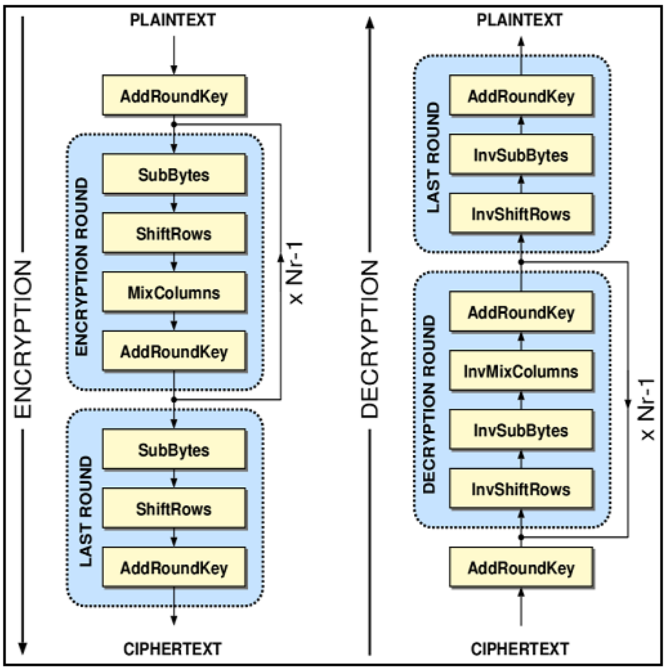

Figure 1: AES Algorithm Structure

Add Round Key: Initial operation is start with add round key operation, where input state array of 16 bytes are XOR with 16 byte portion of key. In the next round key will never be reuse and expansion is perform using key expansion technique. Decryption process is reversed state array is XOR with last 16 bytes portion of the expanded key.

Sub Byte Operation: Each byte portion of the state array is replaced with the other byte according to look up table that is called S-Box. For decryption each value is replace with corresponding inverse of S-Box.
Shift Row Transformation: This operation perform circular shift of last three rows of state array, where second row shift by one byte position to left of the matrix, third row shift by two bytes and forth row shift by three bytes position left in matrix. For decryption process right circular shift will be done.

Mix Column Operation: Each column in state array is multiply by a fixed constant matrix and result stored in same column of the state array.

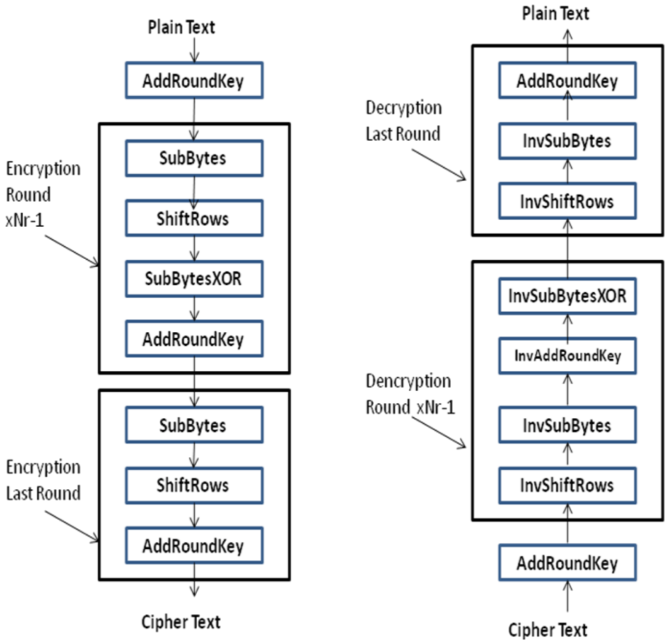

Figure 2: Modify AES Algorithm Structure [12]

Original AES takes too much time to encrypt and decrypt the data. Mix column operation provides more security against different attacks because of its complex computation and that required more computational resources compare to other operations. Modified AES will replace the mix column function with other sub byte XOR function. This modified AES concepts was taken from [12] and it become light weight process that increase speed performance algorithm. Modify AES contains two S-Boxes, first S-Box is the original S-box and second S-Box (new S-Box) will be constructed and replace the mix column operation. And it will be constructed using XOR operation and affine transformation. Modify AES algorithm follows the sequence of sub bytes, shift rows, sub byte XOR and add round key operations for nine rounds. And in the last round sub byte, shift rows and add round key operation will be performed to produce the cipher text. For decryption process sequence of inverse operation is performed that are inverse shift rows, inverse sub bytes, inverse add round and inverse sub byte XOR. Figure 2 show the modified AES algorithm structure.

New S-Box Construction: To construct the new subBytesXOR function in modified AES, Felicisimo V. Wenceslao, Jr apply exclusive OR operation and affine transformation operation for new S-Box. In the first step apply XOR operation on each cell of original AES S-Box and some key[i]. Key[i] shall be any hexadecimal value between 00 to FF. Now this will construct the intermediate part of new S-Box, that is refer as AES-SboxXOR. After constructing the initial values of AES-SboxXOR, apply affine transformation to each cell of AES-SboxXOR. At the end of the affine transformation the final values of the new S-Box are known. Now in subBytesXOR function new generated S-Box is used to replace each byte portion of the state array. This function is same as sub byte operation and it will use new S-Box to replace the each byte of state array. Compare to mix column operation subByteXOR function consume less computation resources in software implementation, so this will increase the encryption and decryption speed. To measure the performance of modified AES experiment conducted on 6400 records on each sites result shown in figure 3.

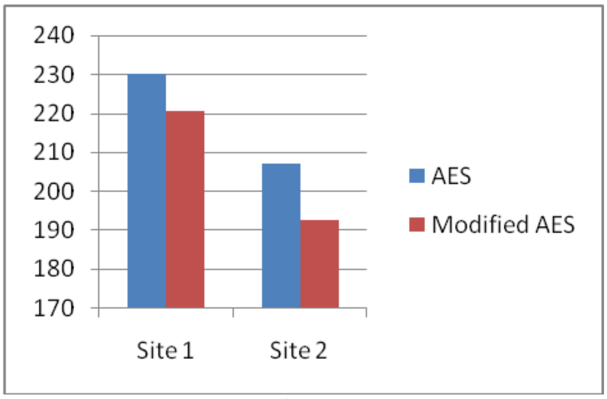

Figure 3: Comparison of Encryption Time in Second

IV. Distributed Privacy Preservation

Distributed privacy preservation algorithm applied to C4.5 decision tree over horizontally partition data. In horizontally fragmentation different sites contain different set of records with the same set of attributes. C4.5 decision tee constructed based on best splitting attributes. Many solutions provide for privacy preservation in horizontal partition data. And the main objective of this proposed solution is to prevent the other party from unauthorized access of data. Proposed solution build a global classification tree among horizontally distributed dataset and not disclosed their private sensitive data among different parties. Figure 3 shows the proposed system flow diagram.

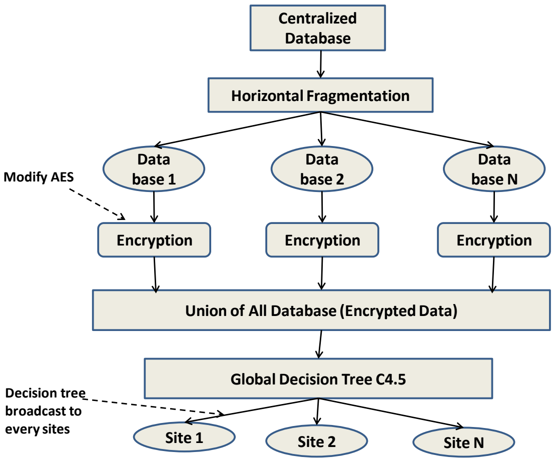

Figure 4: Privacy Preserveing Decision Tree Flow Diagram

As per the flow diagram figure 4 we are first creating the distributed environment, apply horizontal fragmentation and divide data in to multiple sites. To protect the data from unauthorized access we are encrypted using modified AES. There exists one trusted third party server that collects all encrypted data and data is in encrypted form so server cannot identify the data records coming from which sites. So here privacy is preserved and data is not shared between multiple parties. After merging the all records global decision tree is constructed on union of the database. Before the construction of decision tree we are decrypting the all merging records because decision tree cannot constructed on encrypted data. After creating the decision tree it will broadcast to every site. Now individual sites can classify the new test instance no need to communicate with other sites. Only one single decision tree constructed in privacy preserve manner so it reduces the time complexity and also improves the classification accuracy. Error based pruning is performed so it increase the classification accuracy. As considering the drawback of this approach that is cost. Merging of all encrypted data at trusted party server increases the cost but it is only one time cost after construction of decision tree it will be broadcast to the all sites. Here we are only considering the privacy issues rather than cost.

V. CONCLUSION

C4.5 decision tree classification is more suitable for distributed privacy preserving classification and that improve the classification accuracy. Main efforts of the purposed system to enhance the security of the using cryptography block cipher AES. This ensures privacy protection as the data sets are encrypted before they are sent to third parties preventing inadvertent disclosure or theft. This technique hides the whole dataset and all encrypted data merge at trusted third party server, so no information are disclosed between parties. This approach reduces the time complexity because it construct only single decision tree among multiple parties using trusted third party and modified AES that reduce the encryption decryption time also. In the scenario where multiple parties want to build a global decision tree classifier without compromise their privacy, this approach reduces the tree construction time and reduce the cost of communication during classification instance time.

REFERENCES

[1] Majid Bashir Malik, M. Asger Ghazi, Rashid Ali, “Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects”,978-0-7695-4872-2/12 $26.00 © 2012 IEEE, DOI 10.1109/ICCCT.2012.15.

[2] Pui K. Fong and Jens H. Weber-Jahnke, “Privacy Preserving Decision Tree Learning Using Unrealized Data Sets”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 2, FEBRUARY 2012.

[3] Yehuda Lindell, Benny Pinkas, “Privacy Preserving Data Mining”, J. Cryptology (2002) 15: 177–206, DOI: 10.1007/s00145-001-0019-2

[4] Ming-Jun Xiao, Liu-Sheng Huang, Hong Shen, Yong-Long Luo, “Privacy Preserving ID3 Algorithm over Horizontally Partitioned Data”, Proceedings of the Sixth International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT’05)
0-7695-2405-2/05 $20.00 © 2005 IEEE.

[5] Jaideep Vaidya, Chris Clifton, Murat Kantarcioglu, A. Scott Patterson, “Privacy-Preserving Decision Trees over Vertically Partitioned Data”, ACM Transactions on Knowledge Discovery from Data, Vol. 2, No. 3, Article 14, Publication date: October 2008.

[6] Saeed Samet, Ali Miri, “Privacy Preserving ID3 using Gini Index over Horizontally Partitioned Data”, DOI: 10.1109/AICCSA.2008.4493598 · Source: IEEE Xplore.

[7] Mohammad Ali Kadampur, Somayajulu D.V.L.N.,” A Noise Addition Scheme in Decision Tree for Privacy Preserving Data Mining”, JOURNAL OF COMPUTING, VOLUME 2, ISSUE 1, JANUARY 2010, ISSN 2151-9617.

[8] Jayanti Dansana, Debadutta Dey, Raghvendra Kumar, A Novel Approach: CART Algorithm for Vertically Partitioned Database in Multi-Party Environment”, Proceedings of 2013 IEEE Conference on Information and Communication Technologies (ICT 2013).

[9] G. Nageswara Rao, M. Sweta Harini, and Ch. Ravi Kishore, “A Cryptographic Privacy Preserving Approach over Classification”, DOI: 10.1007/978-3-319-03095-1_53, © Springer International Publishing Switzerland 2014.

[10] Gurjeevan Singh, Ashwani Singla, K S Sandha, “Cryptography Algorithm Comparison For Security Enhancement In Wireless Intrusion Detection System", International Journal of Multidisciplinary Research, Vol.1 Issue 4, August 2011, ISSN 2231 5780.

[11] Jaideep Vaidya,Basit Shafiq, Wei Fan, Danish Mehmood, and David Lorenzi, “A Random Decision Tree Framework for Privacy-Preserving Data Mining”, IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, VOL. 11, NO. 5, SEPTEMBER/OCTOBER 2014.

[12] Felicisimo V. Wenceslao, Jr.,” Performance Efficiency of Modified AES Algorithm Using Multiple S-Boxes”, International Journal of New Computer Architectures and their Applications (IJNCAA) 5(1): 1-9 The Society of Digital Information and Wireless Communications, 2015 (ISSN: 2220-9085).