CLASSIFICATION ON HORIZONTAL PARTITION DATA PRIVACY PROTECTING DECISION TREE

N.Sengamalaselvi, Asst.Prof

Dept of Computer Science,

Auxilium College of Arts and Science for Women,

Regunathapuram.

Abstract: Protecting individual data privacy is a big challenge in distributed environments. To create a global decision classifier between multiple parties, it is important to share such data that emerge from privacy concerns with each other. The decision tree privacy preservation algorithm solves this privacy problem in a distributed setting that generates a global classification tree.Various parties. This study proposed a C4.5 privacy protection approach that addresses both discrete and continuous attribute values and uses the Advanced Encryption Standard Protocol to preserve privacy. The algorithm's encryption/ decryption speed is also significant. Modified AES is used to lower the encryption/decryption speed.

Keywords: Data mining, Decision tree, Distributed database, Conservation of privacy.

I.INTRODUCTION

Data mining involves extracting secret information from a vast number of databases, and privacy is the key problem when we deal with the mining process. Data mining privacy protection safeguards personal information in the database and also preserves the utility of data. Without compromising the utility of the data, we need to protect the privacy of the data.Whatever information we are going to safeguard, that produces the same result as normal data. Thus, the primary objective of data mining privacy protection is to reduce the risk of data abuse and at the same time achieve the same results as those obtained in the absence of such privacy preservation techniques [1]. At various stages of the data mining process, we can apply privacy protection techniques.We may apply privacy techniques from the data collection process to the generation of information phase.

The most significant role in data mining is classification, which predicts the class label of a previously unknown case.Here, we concentrate on the classifier of the decision tree, which is one of the most common classification techniques. This follows the supervised learning approach and builds a classification tree based on the training dataset and then adds test attributes for test data classification[2]. The decision tree is the structure of the tree type and its leaf node is the outcome of the classification or class mark,different decision trees ID3, C4.5, CART use various metrics of collection of splitting attributes such as information gain, gain ratio, and gini index respectively. ID3 only addresses the discrete value and does not manage the missing value, and there is no pruningperformed, while C4.5 is able to manage discrete and continuous value and pruning is performed based on costs. CART has all characteristics such as C4.5 and error-based pruning is done, but if any changes in training data occur and it splits only by one variable, CART can become unstable.The Random Decision Tree classifier randomly selects the variable that does not predict the most predictive attribute values of the class label to create the tree. Among all these, applied privacy protection strategies based on the source of input data. If information is stored in a single machine/location, then it is a centralized database environment. And if data is distributed between various parties/machines, then the distributed environment is distributed C4.5 decision tree classifier was used in our proposed scheme.Fragmentation of data in the distributed database is implemented and column wise fragmented data is stored in the vertical partition database and raw wise fragmented data in the horizontal partition database. When we deal jointly with distributed world, data mining tasks are performed and at this point SMC is most suitable. We need to share some information that needs to be protected in time for individual data privacy to create the global decision tree classifier between multiple parties. In order to easily take care of the privacy between various parties, current work uses the SMC methods. We use the AES algorithm here to deal with the protection of privacy.The speed of encryption and decryption is also a significant factor in the privacy protection algorithm. It takes more time for the AES algorithm to encrypt and decrypt the data, so we use a modified AES algorithm that reduces the time for encryption and decryption.

2.TECHNIQUES OF PRIVACY PRESERVATION

The privacy preservation approach takes into account specific parameters, such as cost, complexity, utility, efficiency, protection, etc. Designing an algorithm that meets all the relevant parameters is very difficult. The security improvement parameter is adopted by this paper at any expense using AES. The SMC (Secure Multiparty Computation) protocol was used for all these current works.Where there are more than two parties involved, we use the SMC protocol. The basic concept behind the SMC is that each party only knows its own input and outcome after the secure measurement of the mechanism.

Two adversarial models operate on SMC: (1) Semi-Honest model (2) Malicious model.Semi-honest model which follows the specification of the protocol and tries to obtain additional information from analyzing the messages obtained during the execution of the protocol. The specification of the protocol is not followed in the Malicious model. The model for semi-honest adversaries is simpler to build than for malicious adversaries.On already encrypted data, homomorphic encryption can be performed and the same result can be obtained as the original data. Techniques for CryptographyProviding stable results and less leakage of privacy, but increasing the overhead of encryption and decryption. And because of additional communication overhead, it becomes less effective with larger datasets and requires more parties. Cryptography methods are divided into two sections:The asymmetric key (public key) algorithm and the symmetric key (private key) algorithm are part of the symmetric key and the block cipher and stream cipher.

In asymmetric key encryption, two keys are used for encryption using a public key and decryption using a private key.Block cipher is part of the encryption of symmetric keys that encrypts the block data rather than a bit at a time. We can use block cipher that provides more protection, but with some extra cost, if we are more concerned about privacy. The suggested solution used the AES algorithm to encrypt data that first encrypts data at various locations and transfers encrypted data to trusted third parties.Group to produce the classifier of the global decision tree.

The algorithm for distributed privacy preservation applied to the C4.5 decision tree over horizontal partition data. Different sites contain different records with the same collection of attributes for horizontal fragmentation.Founded on the best dividing attributes, C4.5 decision tee. In horizontal partition details, several solutions provide for privacy preservation. And the main purpose of this suggested approach is to avoid unauthorized access to the data by the other party.

The proposed solution builds a worldwide classification tree between horizontally dispersed datasets and does not reveal its private sensitive data to various parties. The proposed system flow diagram is shown in figure 1.

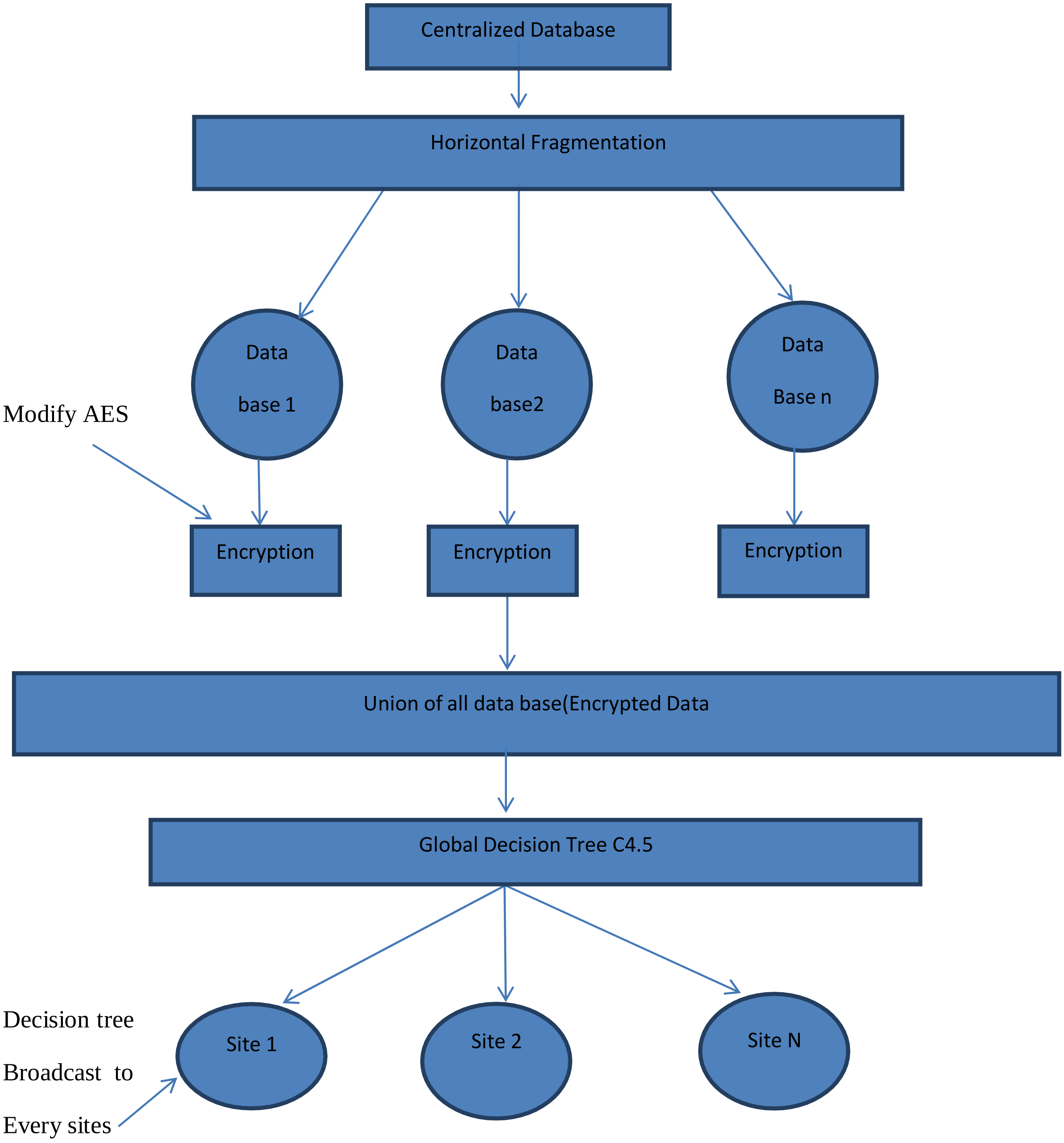
Centralized Database
Horizontal Fragmentation
Data base 1
Data base2
Data Base n
Modify AES
Encryption
Encryption
Encryption
Union of all data base(Encrypted Data
Global Decision Tree C4.5
Decision tree
Broadcast to
Every sites
Site 1
Site 2
Site N

Figure1:Privacy Preserving Decision Tree Flow Diagram

As per Figure 1 of the flow diagram, we first create the distributed environment, apply horizontal fragmentation and divide information into multiple locations. We are encrypted using altered AES to protect the data from unauthorized access.There is a trusted third-party server that collects all encrypted data and encrypted data, so that the server is unable to identify the data records from which sites. So, privacy is maintained here and information is not shared between multiple parties. After merging all records, the global decision tree is built on a database union.We decrypt all merging records before the construction of the decision tree because the decision tree cannot be constructed on encrypted data. It will broadcast to each location after creating the decision tree. Individual sites can now classify the new test instance without any need to interact with other sites.

Only one single decision tree is constructed to preserve privacy in a way that reduces the complexity of time and also improves the accuracy of classification. Error-based pruning is conducted to increase the accuracy of the classification. As the drawback of this approach is considered, that is cost. Merging all encrypted data on a trusted party server increases the cost, but it will be transmitted to all sites only once after construction of the decision tree. We are only considering the issues of privacy here rather than cost.

3. CONCLUSION

The classification of the C4.5 decision tree is more fitting for the classification of distributed privacy preservation and increases the precision of the classification. Key attempts by the intended method to improve the security of the AES cryptography block cipher use.As the data sets are secured until they are submitted to third parties to avoid inadvertent disclosure or theft, this ensures privacy security. This technique conceals the entire dataset and merges all encrypted data on a trustworthy third-party server, so there is nodata is disclosed between parties. This method decreases the time complexity because it uses trusted third parties and updated AES to create only a single decision tree among multiple parties, which also reduces the time of encryption decryption.This method reduces the tree in a situation where multiple parties want to build a global decision tree classifier without violating their privacy construction time and lower communication costs.

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 ANDDATAENGINEERING, VOL. 24, NO. 2, FEBRUARY 2012.

[3] JaideepVaidya, Chris Clifton, Murat Kantarcioglu, A. Scott Patterson, "Privacy-Preserving Decision Trees over Vertically Partitioned Data", ACM Transactions on Knowledge Discovery fromData,Vol.2,No.3,Article14,Publicationdate:October2008.

[4] SaeedSamet, Ali Miri, "Privacy Preserving ID3 using Gini Index overHorizontally Partitioned Data**",** DOI: 10.1109/AICCSA.2008.4493598 Source: IEEEXplore.

[5] G. NageswaraRao, M. SwetaHarini, and Ch. Ravi Kishore, "A Cryptographic Privacy Preserving Approach over Classification**",** DOI: 10.1007/978-3-319-03095-1_53, © Springer International Publishing Switzerland2014.