ISSN(Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Special Issue 6, May 2015

Privacy and Secured Multiparty Data Categorization using Cloud Resources

S.R.Priyadharsani, M.Parthiban, E.Punarselvam

Student, Department of CSE, Sengunthar College of Engineering, Tiruchengode, Namakkal Dt, Tamilnadu, India

Assistant Professor, Department of IT, Muthayammal Engineering College, Rasipuram, Namakkal Dt, Tamilnadu, India.

Assistant Professor, Department of IT, Muthayammal Engineering College, Rasipuram, Namakkal Dt, Tamilnadu, India

ABSTRACT: Data categorization methods are used to assign class labels to the transactional data values. Resource requirement for the data categorization process is very high. In cloud environment users' data are usually processed remotely in unknown machines that users do not own or operate. User data control is reduced on data sharing under remote machines. Anomalous and normal transactions are identified using classification techniques. Neural network techniques are used for the classification process. Back-Propagation Neural network (BPN) is an effective method for learning neural networks. Input layer, hidden layer and output layer are used in the neural network operations.

Shared data values are maintained under different parties to perform the data categorization process. A trusted authority (TA), the participating parties (data owner) and the cloud servers entities are involved in the privacy preserved mining process. TA is only responsible for generating and issuing encryption/decryption keys for all the other parties. Participating party is the data owner uploads the encrypted data for the learning process. Cloud server is used to compute the learning process under cloud resource environment. Each participant first encrypts their private data with the system public key and then uploads the ciphertexts to the cloud. Cloud servers execute most of the operations in the learning process over the ciphertexts. Cloud server returns the encrypted results to the participants. The participants jointly decrypt the results with which they update their respective weights for the BPN network. Boneh, Goh and Nissim (BGN) doubly homomorphic encryption algorithm is used to secure the private data values. Data splitting mechanism is used to protect the intermediate data during the learning process. Random sharing algorithm is applied to randomly split the data without decrypting the actual value. Secure scalar product and addition operations are used in the encryption and decryption process.

The privacy preserved data categorization scheme is composed without the trusted authority for key management process. Key generation and issue operations are carried out in a distributed manner. Cloud server is enhanced to verify the user and data level details. Privacy preserved BPN learning process is tuned with cloud resource allocation process.

I. INTRODUCTION

Cloud computing has achieved tremendous success in offering Infrastructure/Platform/Software as a Service, in an on-demand fashion, to a large number of clients [8]. This is evident in the popularity of cloud software services, e.g., Gmail and Facebook and the rapid development of cloud platforms, e.g., Amazon Elastic Compute Cloud (EC2). The key enabling factor for cloud computing is the virtualization technology, e.g., Xen, that provides an abstraction layer on top of the underlying physical resources and allows multiple operating systems and applications to simultaneously run on the same hardware. As

virtual machine monitors (VMM) encapsulate different applications into each separate guest virtual machine (VM), a cloud provider can leverage VM consolidation and migration to achieve excellent resource utilization and high availability in large data centers.

With the recent adoption and diffusion of the data outsourcing paradigm, where data owners store their data on external servers, there have been increasing general demands and concerns for data confidentiality. Besides well-known risks of confidentiality and privacy breaks, threats to out- sourced data include improper use of information: the server could use substantial parts of a collection of data

gathered and organized by the data owner, potentially harming the data owner's market for any product or service that incorporates that collection of information. Traditional access control architectures assign a crucial role to the reference monitor for ensuring data confidentiality. The reference monitor is the system component responsible of the validation of access requests. The data outsourcing scenario however challenges one of the basic tenets of traditional access control architectures, where a trusted server is in charge of defining and enforcing access control policies. This assumption no longer holds here, because the server does not even have to know the access control policy that is defined by the data owner. We therefore need to rethink the notion of access control in open environments, where external servers take full charge of the management of the outsourced data and are not trusted with respect to the data confidentiality.

An important opportunity for a revision for the access control architecture can be based on the use of cryptography. Cryptography can be considered as a tool that transforms information in a way that its protection depends only on the correct management of a compact secret. Cryptography is typically used when information is transmitted on a channel, with the assumption that the channel lies outside of the trust boundary of the system. The improvements in cryptographic algorithms, extremely reduces the cost for the use of cryptography for stored data, producing a continuous increase in its adoption. A simple application of cryptography to stored resources can then be based on the well-known correspondence between a network and storage service: both organize the information they have to transfer/store in discrete pieces. A more advanced solution takes into account that the nature of the storage service is different. For instance, the authors exploit cryptography to the aim of protecting the sensitive information plaintext represented in memory pages when a trusted process accesses it.

Indeed, the application of cryptography for the protection of files is today available as an option in most modern operating systems, to make it impossible to access the information without access to the keys stored within the system. Encryption reduces the risk of loss of confidential information deriving from low-level access to the devices. The cryptographic protection can also be used to protect the swap area on disk to reduce the risk that processes could access information they are not authorized to see by reading the content of swap pages released by processes managing confidential information. Nonetheless, the cryptography options ordered by current operating systems have been designed to protect local resources and access control is still realized using the services of a reference monitor.

II. RELATED WORK

A considerable amount of methods for privacy preservation in data mining use cryptography techniques from the Secure Multi-party Computation (SMC) area based on the seminal works by Yao and Goldreich et al. Several relevant operations for SMC were defined. These operations are applied, amongst others, in the following data mining algorithms to ensure privacy preservation: Decision tree induction was enhanced for vertically [5] and horizontally partitioned data to generate decision trees without data disclosure. Privacy preserving association rule mining was proposed as well as clustering methods. A summary of data mining applications and their privacy preserving solutions is given by Vaidya et al. [4]. In the area of neural networks, the aspect of privacy preservation is mostly disregarded.

Wan et al. present a generic formulation for secure computation of gradient descent methods [9]. The authors discuss a multi-party-protocol for vertically partitioned data that can be used to train a neural network. To ensure privacy, the target function is defined as a composition of two functions. Thus, the weights can be adapted locally. A second protocol for a secure summation of two scalar products is also suggested as a part of the overall process.

A privacy preserving version of self organizing maps (SOM) is presented by Han et al. [10]. SOMs belong to the class of unsupervised learning techniques and are applied, e.g., for dimension reduction. The authors present a two-party protocol to adapt the network weights iteratively for vertically partitioned data. Barni et al. address in [3] a two-party privacy preserving protocol for neural network based computation. In their setting, the first party owns the confidential data and the second party owns the confidential model that is applied on the first party's data. Both, the data and the network model, are kept private. The approach does not consider how the used network model is actually trained but assumes that it already exists.

III. PRIVACY PRESERVED DATA MINING CONCEPTS

Data mining and knowledge discovery in databases are two new research areas that investigate the automatic extraction of previously unknown patterns from large amounts of data. Recent advances in data collection, data dissemination and related technologies have inaugurated a new era of research where existing data mining algorithms should be reconsidered from a different point of view, this of privacy preservation. It is well documented that this new without limits explosion of new information through the Internet and other media, has reached to a point where threats against the privacy are very common on a daily basis and they deserve serious thinking.

Privacy preserving data mining is a novel research direction in data mining and statistical databases, where data mining algorithms are analyzed for the side-effects they incur in data privacy. The main consideration in privacy preserving data mining is twofold. First, sensitive raw data like identifiers, names, addresses and the like should be modified or trimmed out from the original database, in order for the recipient of the data not to be able to compromise another person’s privacy. Second, sensitive knowledge which can be mined from a database by using data mining algorithms should also be excluded, because such knowledge can equally well compromise data privacy. The main objective in privacy preserving data mining is to develop algorithms for modifying the original data in some way, so that the private data and private knowledge remain private even after the mining process. The problem that arises when confidential information can be derived from released data by unauthorized users is also commonly called the “database inference” problem.

IV. CLOUD DATA MINING WITH PRIVACY

Back-Propagation is an effective method for learning neural networks and has been widely used in various applications. The accuracy of the learning result, despite other facts, is highly affected by the volume of high-quality data used for learning. As compared to learning with only local data set, collaborative learning improves the learning accuracy by incorporating more data sets into the learning process: the participating parties carry out learning not only on their own data sets, but also on others’ data sets. With the recent remarkable growth of new computing infrastructures such as cloud computing, it has been more convenient than ever for users across the Internet, who may not even know each other, to conduct joint/collaborative learning through the shared infrastructure [11].

In this work, we address this open problem by incorporating the computing power of the cloud [7]. The main idea of our scheme can be summarized as follows: each participant first encrypts her/his private data with the system public key and then uploads the ciphertexts to the cloud; cloud servers then execute most of the operations pertaining to the learning process over the ciphertexts and return the encrypted results to the participants; the participants jointly decrypt the results with which they update their respective weights for the BPN network. During this process, cloud servers learn no privacy data of a participant even if they collude with all the rest participants. Through offloading the computation tasks to the resource-abundant cloud, our scheme makes the computation and communication complexity on each participant independent to the number of participants and is, thus, highly scalable. For privacy preservation, we decompose most of the sub algorithms of BPN network into simple operations such as addition, multiplication and scalar product.

To support these operations over ciphertexts, we adopt the Boneh, Goh and Nissim (BGN) “doubly homomorphic” encryption algorithm and tailor it to split the decryption capability among multiple participants for collusion-resistance decryption.

As decryption is limited to small numbers, we introduce a novel design in our scheme such that arbitrarily large numbers can be efficiently decrypted. To protect the intermediate data during the learning process, we introduce a novel random sharing algorithm to randomly split the data without decrypting the actual value. Thorough security analysis shows that our proposed scheme is secure. Experiments conducted on Amazon Elastic Compute Cloud (Amazon EC2), over real data sets from UCI machine learning repository, show that our scheme significantly outperforms existing ones in computation/communication cost and accuracy loss.

V. NEURAL NETWORK LEARNING UNDER MULTIPARTY ENVIRONMENT

In this paper, we aim at enabling multiple parties to jointly conduct BPN network learning without revealing their private data. The input data sets owned by the parties can be arbitrarily partitioned. The computational and communicational costs on each party shall be practically efficient and the system shall be scalable. Specifically, we consider a 3-layer neural network for simplicity but it can be easily extended to multilayer neural networks. The learning data set for the neural network, which has N samples, is arbitrary partitioned into $Z(Z \geq 2)$ subsets. Each party Ps holds

$x_1{}^m$, $x_2{}^m$,...., $x_a{}^m$ and has

x^m x^m .... x^m x^m
1 1
1 2 1Z 1
……
$_x m$ x^m .... x^m x^m
a $a1$ $a1$
1 2 Z $a.1$

Each attribute in sample x^m , x^m ,...., x^m ,1 m N, is possessed by
$\langle$ 1 2 a $\rangle$
only one party—if Ps possesses x^m ,1 k a,
k
then x^m x^m otherwise x^m 0 . In this paper,
ks k ks
$w^h{}_{jk}$ denotes the weight used to connect the input layer node k and the hidden layer node j; $w_{ij}{}^o$

denotes the weight used to connect the hidden layer node j and the output layer node i, where 1 k a, 1 j b, 1 i c and a, b, c are the number of nodes of each layer [1]. For collaborative learning, the main task for all the parties is to jointly execute the operations defined in the Feed Forward stage and the Back-Propagation stage as shown in Algorithm 1. During each learning stage, except for the final learned network, neither the input data of each party nor the intermediate results generated can be revealed to anybody other than TA.

$\longrightarrow$

Input : N input sample vectors V i ,1
i N with a dimensions
iteration $_{\max}$,learning rate ,target value
t $_i$ sigmoid function $\frac{1}{\quad}$
$f\,x$ 1 $_e$ x
Output : Network with final weights :
$_w h$, w ,1 k a,1 j b,1 i C
j i,
k j
1 begin

```
Randomly Initiallize all $w^{h}_{jk}$ , $w_{ij}^{o}$.
  for iteration = 1,2…..,iteration $_{max}$ do
    for sample = 1,2…, N do // Feed
Forward Stag :
      for j = 1,2…, b do
        $h_j \quad f \quad \underset{k\ 1}{a} \quad x_k * w^{h}_{jk}$
      for j = 1,2…, c do
        $o_i \quad f \quad \underset{j\ 1}{a} \quad h_{jk} * w_{ij}^{o}$
      if Error=$\frac{1}{2}$ $_{i}^{c}{}_{1} (t_i \quad o_i)^2$ threshold
then // Back – Propagation Stage :
        $w_{i,j}^{o} \quad (t_i \quad o_i) * h_j$
        $w_{jk}^{h} \quad h_j\ 1\ h_j\ x_k\ {}_{i}^{c}{}_{1}[(t_i \quad o_i) * w_{ij}^{o}]$
        $w_{ij} \quad w_{ij}\ w_{ij}$
        $w_{ik}^{h} \quad w_{ik}^{h}\ w^{h}_{jk}$
      else // Learning Finish
        break
```

Algorithm 1: Back-Propagation Neural Network Learning Algorithm

To achieve the above goals, the main idea of our proposed scheme is to implement a privacy preserving equivalence for each step of the original BPN network learning algorithm described in Algorithm 1. Different from the original BPN network learning algorithm, our proposed scheme lets each party encrypt her/ his input data set and upload the encrypted data to the cloud, allowing the cloud servers to perform

most of the operations, i.e., additions and scalar products. To support these operations over ciphertexts, we adopt and tailor the BGN "doubly homomorphic" encryption for data encryption [2]. Nevertheless, as the BGN algorithm just supports one step multiplication over ciphertext, the intermediate results shall be first securely decrypted and then encrypted to support consecutive multiplication operations as described in Algorithm 1. For privacy preservation, however, the decrypted results known to each party cannot be the actual intermediate values, for example, the values of the hidden layer. For this purpose, we design a secret sharing algorithm that allows the parties to decrypt only the random shares of the intermediate values. The random shares allow the parties to collaboratively execute the following steps without knowing the actual intermediate values [6]. Data privacy is thus well protected. The overall algorithm is the privacy preserving equivalence of Algorithm 1. We propose three other cloud-based algorithms for secure scalar product and addition, secure random sharing and sigmoid function approximation process. After the entire process of the privacy preserving learning, all the parties jointly establish a neural network representing the whole data set without disclosing any private data to each other.

VI. ISSUES ON PRIVACY PRESERVED DATA CATEGORIZATION SCHEMES

Collaborative BPN network learning is applied over arbitrarily partitioned data. A trusted authority (TA), the participating parties (data owner) and the cloud servers entities are involved in the privacy preserved mining process. TA is only responsible for generating and issuing encryption/decryption keys for all the other parties. Participating party is the data owner uploads the encrypted data for the learning process. Cloud server is used to compute the learning process under cloud resource environment. Each participant first encrypts their private data with the system public key and then uploads the ciphertexts to the cloud. Cloud servers executes most of the operations in the learning process over the ciphertexts. Cloud server returns the encrypted results to the participants. The participants jointly decrypt the results with which they update their respective weights for the BPN network. Boneh, Goh and Nissim (BGN) doubly homomorphic encryption algorithm is used to secure the private data values. Data splitting mechanism is used to protect the intermediate data during the learning process. Random sharing algorithm is applied to randomly split the data without decrypting the actual value. Secure scalar product and addition operations are used in the encryption and decryption process. The following issues are identified from the current cloud data categorization schemes.

- Centralized key distribution model
- Malicious party attacks are not handled
- Noisy data upload is not controlled

- Resource allocation and data distribution is not optimized

VII. SECURED MULTIPARTY DATA CATEGORIZATION SCHEME

The collaborative learning process is handled without the Trusted Authority (TA). Key generation and issue operations are carried out in a distributed manner. Cloud server is enhanced to verify the user and data level details. Privacy preserved BPN learning process is tuned with cloud resource allocation process. The cloud data analysis process is designed to utilize the cloud resources for the training process. Key aggregation process is used to generate and share the key values. Training is performed under the cloud server with privacy. The system is divided into six major modules. They are cloud server, trusted authority, data provider, upload process, training process and data classification.

The cloud server module is designed to provide resources for the clients. Trusted authority module is designed to manage key distribution process. Data provider is designed to share the data in the cloud. Data encryption and upload process are managed under upload process module. Neural network learning process is carried out under the training process module. Data classification module is designed to classify the client data values.

7.1. Cloud Server

The cloud server manages the user and resource details. User authentication is performed in the cloud server. The cloud server collects resources from different resource providers. Resource scheduling process is used to allocate computational resources to the training process.

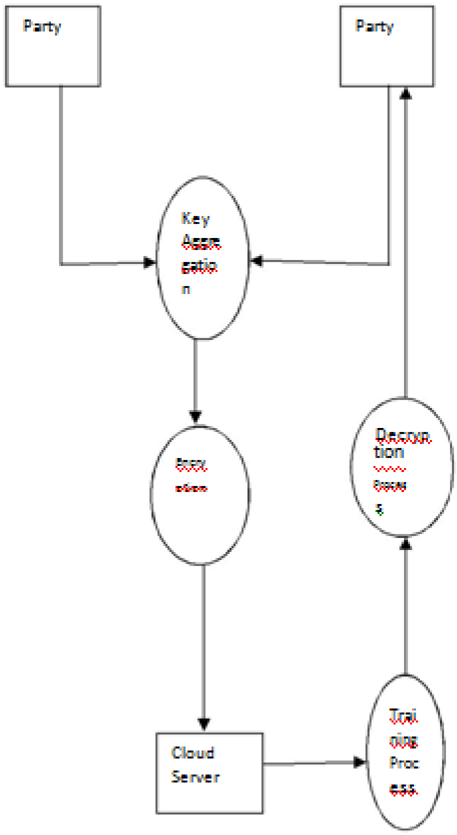

Fig. No: 7.1. Secured Multiparty Data

Categorization Scheme

7.2. Trusted Authority

Trusted Authority (TA) application is used for key management process. Public key and secret key values are generated in the trusted authority. Key values are issued to the data providers. User accounts are verified with cloud server environment.

7.3. Data Provider

Data provider maintains the shared data values. Noise removal process is applied on the data values. Multiple data providers are involved in the data classification process. Data providers are referred as data owner or parties.

7.4. Upload Process

Shared data values are uploaded from the data provider to the cloud server. Encryption process is carried out to secure the sensitive attributes. Boneh, Goh and Nissim (BGN) doubly homomorphic algorithm is used for the encryption process. The data provider uses two types of key generation models. They are Trusted Authority (TA) based key model and Distributed key model. Trusted Authority generates and issues the key value to the data provider. Aggregation based key generation mechanism is used in distributed key model. Labeled transaction data values are collected and updated in the cloud server.

7.5. Training Process

Resource scheduling process is initiated in the cloud server for the training process. Back Propagation Neural network (BPN) algorithm is used for the training process. Random sharing algorithm is used in the data splitting process to secure the intermediate data values. Training process results are redirected to the data provider.

7.6. Data Classification

Trained data values are collected from the cloud server. Data provider decrypts the trained data values. Data encryption/decryption tasks are carried out using secure scalar product and addition mechanism. Test data values are compared with the trained data values for the class assignment process.

VIII. CONCLUSION

The data categorization scheme is adapted for the cloud environment with privacy preserved data sharing mechanism. Data privacy is ensured with encrypted data learning process using cloud resources. Privacy preserved BPN

learning scheme is enhanced without using the Trusted Authority for key management process. The system also handles the malicious party attacks in the learning process. Collaborative learning model improves the classification accuracy level. The system reduces the computational and communication cost in privacy preserved data classification process. Data privacy is improved in all parties. Key generation and issue load is minimized in the aggregation based cryptographic model.

REFERENCES

[1] "The Health Insurance Portability and Accountability Act of Privacy and Security Rules," http://www.hhs.gov/ocr/privacy, 2013.

[2] "National Standards to Protect the Privacy of Personal Health Information," http://www.hhs.gov/ocr/hipaa/finalreg.html, 2013.

[3] Barni, M., Orlandi and C., Piva, A.: A privacy-preserving protocol for neural-network-based computation. In: MM&Sec '06: Proceeding of the 8th workshop on Multimedia and security, New York, NY, USA, ACM Press 2006.

[4] Vaidya, J., Clifton, C., Zhu, M.: Privacy Preserving Data Mining. Volume 19 of Advances in Information Security. Springer, 2006.

[5] Vaidya, J., Clifton, C.: Privacy-preserving decision trees over vertically partitioned data. In: The 19th Annual IFIP WG 11.3 Working Conference on Data and Applications Security, Storrs, Connecticut, Springer, 08 2005.

[6] T. Chen and S. Zhong, "Privacy-Preserving Backpropagation Neural Network Learning," IEEE Trans. Neural Network, vol. 20, no. 10, pp. 1554-1564, Oct. 2009.

[7] Qin Liu, Chiu C. Tan, Jie Wu and Guojun Wang, "Towards Differential Query Services in Cost-Efficient Clouds" IEEE Transactions On Parallel And Distributed Systems, Vol. 25, No. 6, June 2014

[8] Ron C. Chiang and H. Howie Huang, "TRACON- Interference-Aware Scheduling for Data-Intensive Applications in Virtualized Environments" IEEE Transactions On Parallel And Distributed Systems, Vol. 25, No. 5, May 2014

[9] Wan, L., Ng, W.K., Han and S., Lee, V.C.S.: Privacy-preservation for gradient descent methods. In Berkhin, P., Caruana, R., Wu, X., eds.: KDD, New York, ACM Press, 2007.

[10] Han, S., Ng, W.K.: Privacy-preserving self-organizing map. In Song, I.Y., Eder, J., Nguyen, T.M., eds.: DaWaK. Volume 4654 of Lecture Notes in Computer Science., Springer, 2007.

[11] R. Grossman and Y. Gu, "Data Mining Using High Performance Data Clouds: Experimental Studies Using Sector and Sphere," Proc. 14th ACM Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '08), pp. 920-927, 2008.