Classification of Multiparty Outsourced Data with Privacy Preservation

Avinash Thube
Department of Computer Engineering, SavitribaiPhule Pune University, Pune, India, JSPM's RajarshiShahu College of Engineering, Tathawade, Pune

Aniket Patil
Department of Computer Engineering, SavitribaiPhule Pune University, Pune, India, JSPM's RajarshiShahu College of Engineering, Tathawade, Pune

Vinayak Shinde
Department of Computer Engineering, SavitribaiPhule Pune University, Pune, India, JSPM's RajarshiShahu College of Engineering, Tathawade, Pune

ABSTRACT

Back-propagation Neural Network is one of the important machines learning technique for classification. Accuracy of any machine learning technique is improved with volume of datasets. Recent developments in computer networks provide such environment that multiple users connect with each other. This scenario leads in development of machine learning technique with collaborative or joint participation of these multiple users. In collaborative learning multiple parties participate jointly in learning where data is shared by users may contain sensitive information such as corporate data, health care data and personal data. There are chances of data leakage by any corrupt party or intruders. Users are concern about privacy of their datasets due to this main hurdle in collaborative machine learning technique many users are not interested to participate. In this project we are giving solution to privacy of individual users data by converting data in the form of cipher texts where users can use these cipher texts directly in the learning process. A machine learning system working on cipher texts needs various computations. Recent development in cloud computing provides good computing environment where we are utilizing these computations. In this way a system will be developed with neural network machine learning technique working on cipher text and utilizing cloud computing services.

Keywords

Cloud, learning, neural network, back-propagation, privacy preserving, data classification.

1. INTRODUCTION

Most of the companies and even individuals have stated outsourcing their data to cloud due to financial betterment and services utilization. Some of the features of cloud are beneficial and attracts the users. With the increasing utilization of cloud and its various services, security of cloud is also becoming the important concern. For example, if medical researchers want to apply machine learning to study health care problem, they need to gather the raw data from hospitals and the patients detail must be protected, according to privacy rules in Health Insurance Portability and Accountability (HIPAA), which establishes the regulations for the use of Protected Health Information [1].

In artificial intelligence and machine learning, problem of identifying unknown data sample is a problem of classification the classifier is built using training set of known data samples. To build good classifier there is need of large volume of data samples. Building a classifier is not possible to individuals or small scale organizations. There is only one feasible solution to overcome from this problem is to outsource data.

Outsourcing the data over cloud leads to a problem whether cloud server able to provide data classifications to clients. In this context uploading clients own data samples to the cloud leads to privacy issues since the data processed in the cloud can be access by untrusted third party server. [2][10]

Back-propagation Neural Network (BPNN) is one of the important and widely used in practical application machines learning technique for classification. The Artificial Neural Network (ANN) involves two phases: training phase and testing phase. In the training phase cloud server trains an ANN using labeled data arrangement to obtain the classification parameters. In the testing phase any unlabeled data sample search by the client can be classified and labeled to match class by a cloud server using the classification parameters which are generated during training cycle or phase. To improve the accuracy of classification BPNN comes into the action. If ANN fails the weights are adjusted. BPNN reduces the error rate by comparing accumulated error with error goal. Challenges 1) To reduce the communication cost due to each user's participation 2) To protect each user private data samples and generated result during ANN learning process. 3) Classification in encoded form of data samples. 4) Improve the classification accuracy.

2. RELATD WORK

Many privacy preserving BPNN learning techniques have been proposed recently. N.Schlitter [3][9] proposed a privacy preserving BPNN learning without disclosing their respective private dataset. This technique can't protect the intermediate results, which may also contain sensitive data, during learning process. There are several classification algorithms developed in pattern recognition and machine learning for different application. YogachandranRahulamathavan [4][8] introduce privacy preserving multiclass support vector machine for outsourcing the data classification in cloud require kernel mapping functions for non-linear data sample also uses paillier cryptosystem which is not much secure as compared to AES. There still many improvements required for efficient and scalable solution that supports multiparty BPNN with privacy preservation in the multiparty environment, none of the existing schemes have solved all these challenge at the same time.

Within the scope of this paper, we explain the classification techniques. The security issues are tried to solve with help of encryption techniques. Other privacy requirements are also solved with symmetric key algorithm. The security from the cloud service provider and used medium is also maintained.

The architecture provides the service to implement the classification over the encoded cloud data sample. The concept of privacy and classification helps to increase the efficiency.

3. SYSTEM ARCHITECTURE

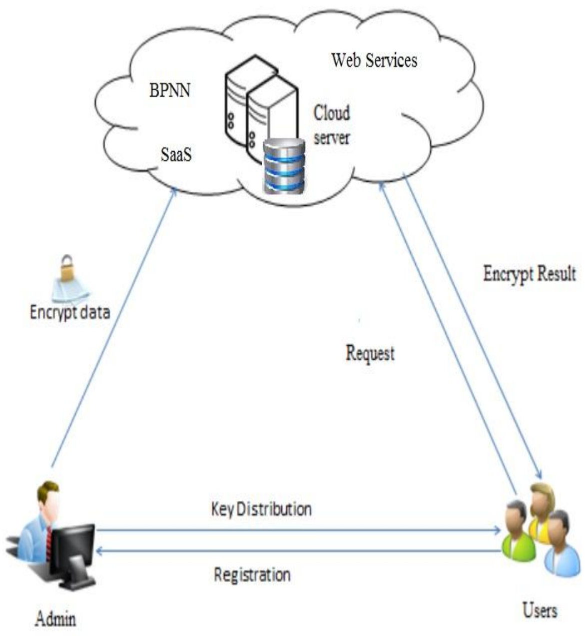

Fig 1: System Architecture

1) In this architecture the main components are user, admin and cloud server.
2) Both Data Owner (Admin) and User should have project software application installed in his/her machine.
3) Admin uploads his data on a server.
4) Application encodes and encrypts the data samples and transfers to cloud; Admin distributes keys to the users who need to access data.
5) Where cloud server decrypts the data, generates ANN parameters and trains network using ANN.
6) User sends his request.
7) Application encodes and encrypts the input samples and sends it to the cloud.
8) Cloud server applies BPNN and generates result then encrypts the result and sends it to the user application.
9) Application decrypts and decodes the result using key.
10) Result is shown to the user.

Input data sample are parsed then encoding of input data sample is perform by using scaling and transformation. By using AES-128 bit key is generated and encryption is performed and encrypted data samples are exported to cloud server.

Cloud-server performs decryption and normalizes the data sample. After normalization phase ANN trains the network. User sends his request to cloud server, server performs prediction using BPNN and generates expected result. To protect generated intermediate results; results are again encrypted using AES and get transfer to the user application. User decrypts and decodes the result using key.

4. PROPOSED SCHEME

4.1 SHA-1

As the name suggests, SHA-1 is a cryptographic function. SHA results a 160-bit (20-byte) hash value. A SHA-1 hash value is rendered as a hexadecimal number, 40 digits long. The previous version of SHA is SHA-0. SHA-1 alters the original hash specification to correct alleged weakness of SHA-0. SHA-1 is the most widely used of the existing SHA hash functions. It is employed in several widely used applications and protocols. SHA-1 algorithm can be used for data integrity and cryptography. The output size for SHA-1 is 160 bits [11]. Internal state size is 160 bits (5*32). Block size is 512 bits. Maximum message size is (2^64)-1 bits. The number of rounds required is 80. Operations performed are and, or, xor, add mod 2^32, rot (rotational) performing the above operations, the algorithm is used to generate the digital signature for the given document[5][13].

4.2 Back-Propagation Neural Network Learning

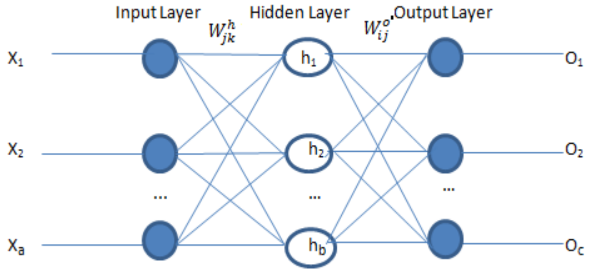

Fig 2: Configuration of Back-Propagation Network

Back-Propagation neural network learning algorithm composed of two stages: feed forward and error back-propagation. [8] Fig. 2 shows a configuration for a three layer (a-b-c) BPN network. We use vector {x_1; x_2; . . . ; x_a} to denote the values of input nodes, vector {h_1; h_2; . . . ; h_b}, and vector {o_1; o_2; . . . ; o_c} to denote the values of hidden layer nodes and output layer nodes, respectively. w_{jk}^{h} is used to represent the weight connecting the input layer node k and the hidden layer node j. w_{ij}^{0} denotes the weight connecting j and the output layer node i, where $1 \leq k \leq a$, $1 \leq j \leq b$, $1 \leq i \leq c$. During the BPN network learning process, the goal is to create a given function by modifying internal weights of input to generate an expected output. All the weights are initialized as small random numbers. In the Feed Forward Stage, values at each layer are computed using the weights, the sigmoid function, and the values at the previous layer. In the Back- Propagation step, the algorithm checks whether the error between output values and target values is within the threshold. If not, all the weights will be modified according to (1), (2) and the learning procedure is repeated. The learning of network will not be terminated until the error is within the threshold or the max number of iterations isexceeded. After the learning, the final weights on each node are used to generate the learned network.

$$\Delta W_{ij}^{0} = -(t_i - o_i\,)h_j$$

$$\Delta W_{jk}^{h} \;=\; -h_j\,(1 - h_{j)}x_k \sum_{i=1}^{c}\big[(t_i - o_i) * W_{ij}^{o}\big]$$

A log-sigmoid function is also known as a logistic function, is given by the relationship

$$\sigma(\text{t}) = 1/(1+(\text{e}\wedge-\beta(\text{t})))$$

Where, ß are a slope parameter. This is called the log-sigmoid because a sigmoid can also be constructed using the hyperbolic tangent function instead of this relation, in which case it would be called a tan-sigmoid. Here, we will refer to the log-sigmoid as simply "sigmoid". The sigmoid function has the property of being similar to the step function.

4.3 AES

AES algorithm is symmetric key algorithm. Symmetric key algorithm uses same key for encryption and decryption. The data is divided into blocks of size 128 bits and can process data blocks of 128 bits using cipher keys of length 128,192 and 256 bits. Depending on the key length, AES is categorized into "AES-128", "AES-192", "AES-256". We have chosen AES as it has only one private key which is distributed among users. Hence it is beneficial for the users to handle single key rather than two keys i.e. public key and private key. It serves the purpose of saving memory space as the data encrypted by AES is of same size as that of original file while, in public key cryptography encrypted file size increases[14].

AES is a comparatively a fast encryption process. It uses less computer resources. Only messages between a particular pair of sender and receiver are affected if a key is compromised. For AES algorithm, the length of cipher key(K) is 128 bits,192 bits and 256 bits, depending on this key length is represented by N_k= 4, 6, or 8 [10]. This N_k reflects the number of 32-bit words (number of columns) in the Cipher Key [10]. For the AES algorithm, the number of rounds to be performed during the execution of the algorithm is dependent on the key size [10]. The number of rounds is represented by N_r, where N_r= 10 when N_k= 4, N_r= 12 when N_k= 6, and N_r= 14 when N_k= 8 [10].

4.4 Scaling

In projective geometry [9], often used in computer graphics, samples are represented using homogeneous coordinates. To scale an object by a vector v = (v_x, v_y, v_z), each homogeneous coordinate vector p = (p_x, p_y, p_z, 1) would need to be multiplied with this projective transformation matrix[9][12].

$$S_V = \begin{bmatrix} Vx & 0 & 0 & 0 \\ 0 & Vy & 0 & 0 \\ 0 & 0 & Vz & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

As shown below, the multiplication will give the expected result

$$S_V P = \begin{bmatrix} Vx & 0 & 0 & 0 \\ 0 & Vy & 0 & 0 \\ 0 & 0 & Vz & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} Px \\ Py \\ Pz \\ 1 \end{bmatrix} = \begin{bmatrix} VxPx \\ VyPy \\ VzPz \\ 1 \end{bmatrix}$$

Since the last component of a homogeneous co-ordinate can be viewed as the denominator of the other three components, a uniform scaling by a common factor s (uniform scaling) can be accomplished by using this scaling matrix

$$S_V P = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & \frac{1}{s} \end{bmatrix}$$

For each vector p = (p_x, p_y, p_z, 1) we would have,

$$S_V P = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & \frac{1}{s} \end{bmatrix} \begin{bmatrix} Px \\ Py \\ Pz \\ 1 \end{bmatrix} = \begin{bmatrix} Px \\ Py \\ Pz \\ \frac{1}{s} \end{bmatrix}$$

This would be homogenized to

$$\begin{bmatrix} SPx \\ SPy \\ SPz \\ 1 \end{bmatrix}$$

5. MATHEMATICAL MODEL

Set Theory

Let G = {U, S, R, L, A, Q, K}

Where,

U is an infinite set of users

S is a cloud server.

R is set of records.

L is finite set of Layers

K is a finite set of keys.

A is a finite set of algorithms

Q is a finite set of queries.

u = Registration(uid, passw, name)

passw = DSA_key(pass)

Already existing user:

Validuser = login(uid, passw)

User(Validuser)

Session starts here.

u = search(query)

Encode(query)

Prediction = BPNN (query)

Server encrypt result = Encryption (result, key)

User decrypts and decodes result using key

Decryption (result, key)

Decode (result, key)

Display (result)

Session ends here.

upload data (data owner as client)

```
Validuser = login(uid, passw)
User(Validuser)
Session starts here.
Parse data = load dataset (dataset.csv)
Transform data = Encode dataset (fmkey, scaling key)
Encrypt data = encryption (key, encoded data)
Send to cloud server
Cloud server decrypts data = decrypt (key , encrypted data)
Normalize data = normalize (encoded data)
Train ANN = ANN (normalized data)
Distribute_keys()     key distribution to the valid users only
Session ends here.
```

6. CONCLUSION

We are integrating symmetric encryption system to provide security to cloud storage, generated results. The problems of security and integrity have been discussed. We purposed the secure, scalable and multiparty back-propagation neural network learning to solve the problem of classification by performing operations over encoded data. The parties encode and encrypt their data and upload the cipher texts to the clouds. The cloud can execute most operation pertaining to the back propagation neural network learning algorithm without knowing any sensitive information. Cost of each party in our scheme is independent to the number of participating parties. We provide encryption at both client and server site. Proposed scheme will produce scalable, efficient and secure classification results. In future it is possible to implement using different encryption techniques such as RSA. To improve accuracy of neural network feed forward neural network is also one of the option.

7. REFERENCES

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

[2] S. Pearson, Y. Shen, and M. Mowbray, "*A Privacy Manager for Cloud Computing*," Proc. Int'l Conf. Cloud Computing (CloudCom), pp. 90-106, 2009.

[3] N. Schlitter, "*A Protocol for Privacy Preserving Neural Network Learning on Horizontal Partitioned Data,"* Proc. Privacy Statistics in Databases (PSD '08), Sept. 2008.

[4] YogachandranRahulamathavan, Suresh Veluru, Kanapathippillai, Cumanan and MuttukrishnanRajarajan*,"Privacy-Preserving Multi-Class Support Vector Machine for Outsourcing the Data Classification in Cloud*", IEEE transaction on Dependable and Secure Computing, vol. 11, no. 5, September/October 2014

[5] AnakAgungPutriRatna, Ahmad Shaugi, Prima DewiPurnamasari, Muhammad Salman,"*Analysis and comparison of MD5 and SHA-1 algorithm implementation in Simple-O authentication based security system*", IEEE Conference on Quality in Research, 2013

[6] C. Lee Giles, Fellow, IEEE, Christian W. Omlin, and Karvel K. Thornber,"*Equivalence in Knowledge Representation: Automata, Recurrent Neural Networks, and Dynamical Fuzzy Systems*", IEEE, vol. 87, no. 9, september 1999

[7] Ming Li, Shucheng Yu, KuiRen, Wenjing Lou And Y. Thomas Hou*,"Toward Privacy-Assured And SearchableCloud Data Storage Services*", Proc. IEEE Network, July/August 2013

[8] Jiawei Yuan, ShuchngYu,"*Privacy Preserving Back-Propagation Neural Network Learning Made Practical with Cloud Computing*", IEEE Transaction on Parallel and Distributed Systems,vol.25,NO.1, January 2014.

[9] Durand Cutler. "Transformations" (http://groups.csail.mit.edu/graphics/classes/6837/F03/lectures/04_transformations.ppt)(PowerPoint). Massachusetts Institute of Technology. Retrieved 12 September 2008

[10] A. Bansal, T. Chen, and S. Zhong, "*Privacy Preserving Back-Propagation Neural Network Learning over Arbitrarily Parti-tioned Data,*" Neural Computing Applications, vol. 20, no. 1, pp. 143-150, Feb. 2011.

[11] C. R. P. Lippmann, "*An introduction to computing with neural networks*," IEEE Acoust. Speech Signal Process. Mag., vol. 4, no. 2, pp. 4–22, Apr.1987.

[12] D. J. Myers and R. A. Hutchinson, *"Efficient implementation of piecewise linear activation function for digital VLSI neural networks*," Electron.Lett., vol. 25, pp. 1662–1663, 1989.

[13] Kasgar A. K., AgrawalJitendra, SahuSantosh, 2012, "*New Modified 256-bit MD5 Algorithm with SHA Compression Funct ion*", IJCA (0975–8887) Volume 42 (12) , pp47-51.

[14] William Stallings, Cryptography and NetworkSecurity: Priciples and Practice,5th Edit ionPrent ice Hall; 5 edit ion (January 24, 2010)