Study of a Privacy Preserving Logistic Regression Algorithm (PPLRA) For Data Privacy in the Context of Big Data

Qiang Chen¹, ², *, Meiling Deng¹

¹ DongGuan City College, DongGuan, China
² Universiti Teknologi MARA, Shah Alam, Malaysia

*Corresponding author e-mail: 2021726553@isiswa.uitm.edu.my

Abstract. Regression algorithms are commonly used in machine learning. Based on encryption and privacy protection methods, the current key hot technology regression algorithm and the same encryption technology are studied. This paper proposes a PPLAR based algorithm. The correlation between data items is obtained by logistic regression formula. The algorithm is distributed and parallelized on Hadoop platform to improve the computing speed of the cluster while ensuring the average absolute error of the algorithm.

Keywords: Data Privacy, Big Data, Logistic Regression, Homomorphic Encryption.

1. Introduction

The big data is highly reliant on the large computer power and huge storage space of cloud computing. In diverse research areas, cloud computing has been studies such as processing of big data. Big data and cloud computing both are intertwined in such a way that later one influences the analysis of former one. A configurable computing environment provided by cloud computing can be effectively managed and quickly deployed. After dropping the computation of data to the cloud, it is easy and convenient to use computing resources without any limitation of other devices. Despite numerous benefits attached to cloud computing, there are still many challenges in privacy preserving. Cloud computing has been found disclosing quality information related to health, population, economy, business etc. With the start of cloud era, privacy preserving phenomenon in relation to big data is currently receiving a lot of attention from practitioners and academic researchers. While collecting big data, ones come across some sensitive and personal information, which once leaked and shared onward, might bring catastrophic consequences. Therefore, there is vital role of encryption for data which is computed in the cloud.

Besides of this, machine learning cannot directly access encrypted data. on the other side, data privacy cannot be ensured once the decrypting key is handed over to the honest-but-curious cloud service. Hence, it is challenging to cope with encrypted data by machine learning algorithm running on the cloud.

Data security and privacy protection are facing great challenges. In recent years, machine learning has made breakthroughs in vision, natural language processing, healthcare and other fields. With the rapid development of machine learning technology, its security and privacy issues have attracted widespread attention. Traditional machine learning training method is to collect training data for centralized training. However, training data often involves people's privacy, such as medical and health
data, interests, political preferences and so on. The traditional model of directly exposing private data to data collectors for model training can no longer be applied to the social environment where people's awareness of privacy protection is enhanced.

Privacy has a wide range of definitions, but in continuous authentication privacy is defined as the user's control over their own data. The issue of privacy protection in continuous authentication has been raised by m. Jakobsson et al. [1], who suggest that there are three ways to solve the user privacy problem:

1) Remove the unique identification of information;
2) Using a pseudonym;
3) Use aggregated data.

However, none of these three methods can effectively solve the problem of user data leakage. The removal of uniquely identified user information can be combined with other public information to recover the user identity, fixed pseudonyms will not prevent the ability of data link, and the use of aggregated data will reduce the authentication accuracy. Therefore, more effective ways are needed to protect the privacy of user data.

2. related work

Presently, there are two methods of privacy preserving namely randomization-based method and encryption method. The randomization method protects data vy adding noise (Agrawal and Srikant, 2000). Other method uses numerous encryptions to take care of data privacy. Under second category of data privacy method, the proposed PPLRA is useful to cope with big data privacy issue. But, in their case, the aspect of cloud computing is clearly missing in relation to data privacy.

Homomorphic encryption, a privacy preserving regularization logistic regression, is very good at solving the interference optimization problem. Aono [6] has developed privacy preserving logistic regression algorithm by adding cost function into logistic regression to polynomial where data comes from distributed data source. Furthermore, there is also availability of privacy preserving logistic regression in relation to multiparty collaborative training. algorithms ...all these studies have focused on machine learning algorithms, but missing aspect is role of cloud computing. The difference between existing studies and the proposed PPLRA is computation based on cloud in this study.

In this study, a privacy preserving logistic regression algorithm (PPLRA) is developed to deal with issue of data privacy in data computing. This study will use homomorphic encryption to control disclosure of private data in PPLRA. Additionally, this study will perform most of the computing tasks to increase computing efficiency. The Sigmoid function of logistic regression entails division and exponential operations, and homomorphic encryption cannot process these functions directly.

In this regard, the present study will use Taylor’s theorem to approximate privacy Sigmoid function through a polynomial sum. In the PPLRA, this study will intend to address following problems.

1) Security calculation of different mandatory operations such as multiplication, addition, and nonlinear Sigmoid function in PPLRA. These calculations are viewed as a source to protect data privacy.
2) By deciding about putting calculations either on cloud side or on the local inside, this study sets data privacy preserving with computational efficiency.
3) Homomorphic encryption method to encrypt data: as compared to other methods, homomorphic encryption is useful to perform direct operation in ciphertext and generates results like that of plaintext.
4) The transformation of Sigmoid function: Such transformation enables sigmoid to function with the application of multiplication ad addition operations.

This research will contribute to the literature by presenting a PPLRA for data privacy purpose and enhancing the efficiency by isolating encryption/decryption operation from cloud computing tasks.

3. Basic Theory

3.1. Regression Algorithms

Regression is one of the oldest but very powerful tools in mathematical modeling, classification, and prediction [2]. Regression has applications in engineering, physics, biology, finance, social sciences and
other fields, and is a basic tool commonly used by data scientists. Regression is often the first algorithm used in machine learning. By learning the relationship between dependent variables and independent variables, the data can be predicted. For example, when estimating housing prices, it is necessary to determine the relationship between the area of a house (the independent variable) and its price (the dependent variable). This relationship can be used to predict the price of a house of a given size. There can be multiple independent variables that affect the dependent variable.

Therefore, regression has two important components: the relationship between independent and dependent variables, and the strength of the influence of different independent variables on the dependent variable. The common regression methods include linear regression, logistic regression and regularization.

As a widely used basic machine learning algorithm, linear regression is a method to model the correlation between one or more independent variables and dependent variables by fitting a number of HBS rows of influencing factors and results. In order to improve and optimize the performance of regression model, it is usually necessary to train a large amount of original data, but the limited local computing resources and storage resources make it difficult for some enterprises or organizations to meet the needs of training model alone.

3.2. Privacy Protection Techniques

M. ai-Rubaie [3] represents that in the field of machine learning, privacy protection of user data can be carried out by means of encryption, disturbance or dimension reduction. Encryption methods mainly include homomorphic encryption, confounding circuit and key sharing, perturbation methods mainly include differential privacy and local differential privacy, and dimension reduction methods mainly include PCA (Principal Component Analysis), DCA (Discriminant Component Analysis) and MD (Multidimensional Degree).

The concept of Homomorphic Encryption (HE) was proposed by Rivest et al. [4] in 1978. The biggest feature of HE is that it can directly perform specific algebraic operations on ciphertext data, and the results obtained by the operations are still encrypted. Same as the result of the same operation on plaintext. Therefore, the homomorphic encryption algorithm can be used to encrypt user data, and then the encrypted data is sent to the remote server for computing or storage. During the computing process, the data does not need to be decrypted, thus ensuring the privacy and security of user data.

3.3. PCA dimensionality reduction

Principle Component Analysis (PCA) [5] reduces dimension through orthogonal decomposition, and its core is to select new mutually orthogonal spatial basis vectors to express the original data. PCA dimensionality reduction can map the original user behavior feature vector into a new vector space through spatial transformation, thus generating new user behavior feature vector, so as to protect the privacy of the original user data.

In order to make continuous authentication system with high accuracy and privacy protection features, the most by adopting the combination of machine learning algorithms and privacy protection scheme, the first kind of scheme with the method of homomorphic encryption combined with logistic regression, makes the system have the characteristics of privacy protection and high accuracy, but the system running time consuming is higher. The second scheme combines PCA dimension reduction with One Class SVM, which makes the system have the characteristics of privacy protection and low latency, but the authentication accuracy of the system is slightly lower than the first scheme.
4. Preliminaries

Logistic regression is a general linear regression analysis model. It is also a supervised learning algorithm. It is generally applied to regression, binary classification, and multi-classification. The application of logistic regression comprises of three steps such as identification of a prediction function, development of a loss function, and recognition of regression parameters suitable to reduce a loss function. In order to use logistic regression for classification issue or regression, there is needed to first create cost function. Later, an iterative optimization model is applied to identify optimal model parameters. At the end, quality of model is tested. In this case, prediction function is related to the Sigmoid function which is written in equation form as below;

\[ S(x) = \frac{1}{1 + e^{-x}} \]  

(1)

The prediction function of logistic regression is written in below equation.

\[ g_w(x) = S(w^Tx) \frac{1}{1 + e^{-w^Tx}} \]  

(2)

In this study, the loss function estimates the difference between the real value and the predictive value. In logistic regression, the maximum likelihood method will be sued to identify the parameters of the model. The loss function is written in mathematical form as below;

\[ L(w) = \frac{1}{m} \left[ \sum_{i=1}^{n} (y_i \log g_w(x_i) + (1-y_i)\log(1-g_w(x_i))) \right] \]  

(3)

Rivest[4] has initially introduced the homomorphic encryption. According to homomorphic encryption, there is consistency between original data and processed data such as both encrypted and decrypted data is same. In other way, the operation of homomorphic encryption on both plaintext and ciphertext delivers same results. Later, Rivest et al., has developed the RSA algorithm. Along with these, there are ElGamal algorithm and Paillier algorithm in the literature as well. Yet, all these algorithms cannot support both homomorphic addition and homomorphic multiplication at the same time. Resultantly, Gentry [6] have introduced another algorithm know as fully homomorphic encryption (FHE) which is good for both addition and multiplication operation. Below figure-1 exhibits the homomorphic encryption scheme. This scheme encrypts plain text data into cipher text through public key and the decrypt cipher text data through private key.

![Figure 1. Model of Homomorphic Encryption](image)

Homomorphic encryption algorithms can usually be divided into additive homomorphism and multiplicative homomorphism according to the type of computation in the encryption domain.

It is assumed that x and y respectively represent the plaintext messages to be encrypted, Enc() represents homomorphic encryption function, the operation symbols “+” and “.” respectively represent the addition and multiplication operations of the information on the plaintext, and the operation symbols “+” and “.” respectively represent the addition and multiplication operations of the message on the ciphertext [7].
For a homomorphic encryption operation satisfying formula (4), it is called an additive homomorphism, and satisfying formula (4), it is called a multiplicative homomorphism. So far, not all homomorphic encryption algorithms can support addition and multiplication at the same time. According to the operation types supported by Homomorphic Encryption Algorithms, Homomorphic Encryption can be divided into two categories. The first is Partial Homomorphic Encryption. The second type is Fully Homomorphic Encryption. Partial Homomorphic Encryption only supports one algebraic operation under the ciphertext, and full Homomorphic Encryption can support any operation under the ciphertext, but so far there is no full Homomorphic Encryption Algorithm can be applied in practice.

PCA dimension reduction technology first takes several important principal component vectors as the basis vectors, and then uses the reconstruction method to reduce the dimensions of the original features [10]. To put it simply, PCA dimension reduction is: For a multidimensional eigenvector \([x_1, x_2, ..., x_n]\), obtain a new vector, and then map the eigenvector to the new eigenvector space. For example, if it is about dimensional reduction of \(nD \rightarrow kD\), then the previously mentioned \(n\)-dimensional eigenvector \([x_1, x_2, ..., x_n]\) maps to the new space with the \(k\)-dimensional vector \([w_1, w_2, ..., w_k]\).

Input conditions and output results of PCA dimension reduction algorithm:
Input: sample set \(D = \{x_1, x_2, ..., x_m\}\), contains \(m\) data in total, and the spatial dimension \(k\) after dimension reduction.
Output: reduced dimension matrix \(W = [w_1, w_2, ..., w_k]\).

PCA dimension reduction algorithm steps are as follows:
1) The following formula is used to centralize all samples.

\[
Enc(x) + Enc(y) = Enc(x + y)
\]

\[
Enc(x) \cdot Enc(y) = Enc(x \cdot y)
\]

2) Calculate the covariance matrix \(XX^T\) of the samples.
3) Perform eigenvalue decomposition for the covariance matrix \(XX^T\).
4) Take the eigenvector \([w_1, w_2, ..., w_k]\) corresponding to the largest \(k\) eigenvalues.

5. Experimental Results and Analysis

5.1. Experimental platform and data description
According to the above study, four computers were used as hardware, including one Master node and three Salave nodes. Ubuntu 18.10, Hadoop version 3.1.2, JDK version 1.8.0_151.

The experiment in this paper selects merchant transaction data set from an e-commerce network, and downloads data sets with sizes of 50KB, 10MB and 500MB respectively for experiment. The details of the data set are as follows: 1) 50KB: including 3000 shopping records of 156 goods by 99 users; 2) 10MB: 5210902 shopping records for 9800 items from 1804 users; 3) 500MB: including 23607456 shopping records of 92275 users for 12058 movies. Each user has bank information, account number, identity and other information, and these information are recorded in the transaction process.

5.2. Evaluation index of experimental results
According to the analysis of server load, as shown in Figure 2, with the increase of user data, the server load gradually increases. As users submit more queries, the hit ratio of client cache is low, and PPLRA algorithm performs poorly at all moving speeds. The server-side load decreases as the cache size increases, as shown in Figure 2, because the approach in this article uses cache objects to answer many client queries directly, reducing the server load.
According to the security analysis, when a user sends a query request to the location service provider and the anonymous end caches all the result data, the user directly extracts POIs from the cache. During this process, the user does not interact with the location service provider, and the location service provider cannot obtain any information from the user. If the user cannot retrieve the result data in the cache, the anonymizer forms a hidden region and sends it to the location server. Even though the location server recognizes all users in the hidden area, the hidden area has at least K users, so it can guess the specified user with a probability of only 1/K.

6. Conclusion
With the rapid development of cloud computing, this problem is well solved. At present, many cloud computing service commercial platforms, such as Amazon and Google, allow clients to upload data to another business server, which operates various machine learning tasks. However, because cloud computing is not credible, it can view and record user data, and may even be attacked by the enemy to leak user data, so it is particularly important to study the linear regression scheme that can protect privacy.

The method based on differential privacy implements privacy protection by adding appropriate noise to the regression model, but the introduction of noise will lead to a decline in the model performance.

Homomorphic Encryption (HE) based scheme usually requires the client to use Homomorphic Encryption algorithm to encrypt the training data, and then the cloud server uses Homomorphic property to conduct training on ciphertext. However, using homomorphic encryption algorithm to encrypt a large number of data is too expensive for the client, and because of the limitation of homomorphic encryption algorithm itself, it can not achieve arbitrary addition and multiplication, so it is not practical in the real environment.

References
[1] Aono, Y., Hayashi, T., Le, T.P., Wang, L., 2016. Privacy-preserving logistic regression with distributed data sources via homomorphic encryption. IEICE Trans. Info Syst. E99.D, 2079–2089.
[2] Bonte, C., Vercauteren, F., 2018. Privacy-preserving logistic regression training. BMC Med. Genom. 11, 86.
[3] Du, W., Li, A., Li, Q., 2018. Privacy-preserving multiparty learning for logistic regression. In: International Conference on Security and Privacy in Communication Systems, pp. 549–568.
[4] Fan, Y., Lin, X., Tan, G., Zhang, Y., Dong, W., Lei, J., 2019. One secure data integrity verification scheme for cloud storage. Future Generat. Comput. Syst. 96, 376–385.
[5] Mohassel, P., Zhang, Y., 2017. Secureml: A system for scalable privacy-preserving machine learning. In: 2017 IEEE Symposium on Security and Privacy (SP). IEEE, pp. 19–38.
[6] Wei, L., Zhu, H., Cao, Z., Dong, X., Jia, W., Chen, Y., et al., 2014. Security and privacy for storage and computation in cloud computing. Inf. Sci. 258, 371–386.

[7] Xu, K., Hao, Y., Guo, L., Guo, Y., Fang, Y., 2015. Privacy-preserving machine learning algorithms for big data systems. In: IEEE International Conference on Distributed Computing Systems.

[8] Tan Z, Li K*, Wang Y. Differential Evolution With Adaptive Mutation Strategy Based on Fitness Landscape Analysis [J]. Information Sciences, 2021,549:142-163.

[9] TIAN Y, LI Y, LIU X, et al. PriBioAuth: Privacy-Preserving Biometric-Based Remote User Authentication[C]/2018 IEEE Conference on Dependable and Secure Computing (DSC). 2018: 1–8.

[10] AL-RUBAIE M, CHANG J M. Privacy preserving machine learning: Threats and solutions [J].arXiv preprint arXiv:1804.11238, 2018.