Multiple Linear Regression Based on Stream Homomorphic Encryption Computing

Yi-Zhuo Zhang
College of Mathematics and Computer Science,
Fuzhou University,
Fuzhou City, China
zhangyz86@hotmail.com

Yiwei Liu
College of Mathematics and Computer Science,
Fuzhou University,
Fuzhou City, China
m18950398636@163.com

Chan-Liang Chung
College of Mathematics and Computer Science,
Fuzhou University,
Fuzhou City, China
andre37@alumni.nccu.edu.tw
ORCID: 0000-0001-7548-3607

Chi-Hua Chen
College of Mathematics and Computer Science,
Fuzhou University,
Fuzhou City, China
chihua0826@gmail.com
ORCID: 0000-0001-7668-7425

Feng-Jang Hwang
School of Mathematical and Physical Sciences,
University of Technology Sydney,
Sydney City, Australia
feng-jang.hwang@uts.edu.au
ORCID: 0000-0002-1741-5590

***Abstract*—This study proposes a multiple linear regression architecture based on stream homomorphic encryption computing to analyze ciphertext for massive secure data computing. The proposed architecture contains three subsystems including terminal subsystem, data access subsystem, and data computing subsystem. The method used behind the presented architecture contains four stages which are data preprocessing stage, data access stage, data computing stage, and result processing stage. In the practical experiments, a case study of traffic information prediction was selected to evaluate the proposed system and method. The predicted traffic information was generated by the proposed method in accordance with the encrypted traffic information. Our experimental results showed that the proposed architecture can effectively and promptly obtain the predicted traffic information.**

***Index Terms*—homomorphic encryption, multiple linear regression, real-time streaming data analysis**

I. Introduction

With the privacy application development of the Internet of things, more and more terminal devices access the Internet and transmit data via web services. However, massive calculations are needed by encryption and decryption for processing huge amounts of data, for instance, the time complexities of RSA (Rivest-Shamir-Adleman) [1] and ECC (elliptic curves cryptography) [2] for encryption and decryption are both $O(m)$, where m is the number of records. Therefore, this study proposes a multiple linear regression system coupled with method based on stream homomorphic encryption computing. The proposed system contains three subsystems including terminal sub-system, data access subsystem, and data computing subsystem. The proposed method consists of four stages which are data preprocessing stage, data access stage, data computing stage, and result processing stage. The contributions of this study are listed as follows.

(1). This study combines the multiple linear regression and homomorphic encryption to calculate the ciphertexts of records and obtain the plaintext of results in secure environments. (2). This study designs and implements the architecture of decentralized computing system to assign tasks to several node devices for processing huge amounts of data. (3). This study designs and implements the techniques of decentralized computing to perform homomorphic-encryption-based multiple linear regression as the map and reduce functions in the MapReduce framework for big data processing.

The remainder of the paper is organized as follows. Section II presents a multiple linear regression system based on stream homomorphic encryption computing, and Section III illustrates the proposed multiple linear regression method based on stream homomorphic encryption computing. The practical experimental results and discussions are given in Section IV. Section V concludes the contributions of this study and discusses the future work.

II. The Proposed System

The proposed multiple linear regression system based on stream homomorphic encryption computing contains the following three subsystems (1) terminal subsystem, (2) data access subsystem, and (3) data computing subsystem, as shown in Fig. 1.

A. A Case Study of Traffic Information Prediction

In this subsection, the multiple linear regression based on stream homomorphic encryption computing is used to predict

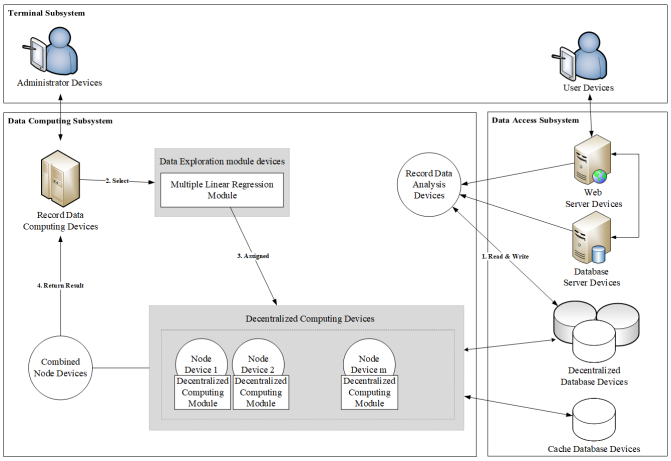

Fig. 1. The proposed multiple linear regression system

traffic information to demonstrate the proposed system.

1) Data Processing Devices: The data processing devices can collect the arrival time from OBUs and measure the travel time between each pair of stations, for instance, the travel time $t^r_{i-n-j,i-n}$ denotes the travel time from the $(i\text{–}n\text{–}j)$-th station to the $(i\text{–}n)$-th station in the r-th record. The data processing devices can save the travel time into decentralized database devices for further analysis and processing.

2) Decentralized Database Servers: In this case, the techniques of HBase and MongoDB are adopted to develop the decentralized database servers and to store the travel time between each two stations.

3) Data Mining Devices: The data mining devices contain the multiple linear regression models (e.g. $T_{t_{i,i-n,i-n-j}}(t^r_{i-n-j,i-n})$) to analyze the correlation of travel time between each pair of stations. This case considers k weighted linear regression models based on m historical records to predict the travel time $t^{r\prime}_{i-n,i}$ from the $(i\text{–}n)$-th station to the i-th station (shown in Equation (1)). The weights of k weighted linear regression models are measured by Equation (2) in accordance with the accuracies of these linear regression models based on m historical records. The slope $a_{t_{i,i-n,i-n-j}}$ and intercept $b_{t_{i,i-n,i-n-j}}$ of the linear regression model $T_{t_{i,i-n,i-n-j}}(t^r_{i-n-j,i-n})$ can be estimated by Equations (3) and (4) according to m historical records. The travel time $t^{r\prime}_{i-n,i}$ can be denoted as Equation (5) in accordance with Equations (1), (2), (3) and (4).

$$t^r_{i-n,i} = \frac{\sum_{j=1}^{k} w_{t_{i,i-n,i-n-j}} \times T_{t_{i,i-n,i-n-j}}(t^r_{i-n-j,i-n})}{\sum_{j=1}^{k} w_{t_{i,i-n,i-n-j}}}. \quad (1)$$

$$w_{t_{i,i-n,i-n-j}} = 1 - \frac{1}{m}\sum_{r=1}^{m} \frac{|t^r_{i-n,i} - t^{r\prime}_{i-n,i}|}{t^r_{i-n,i}}. \quad (2)$$

$$a_{t_{i,i-n,i-n-j}} = \frac{\sum_1}{\sum_2}, \quad (3)$$

where

$$\sum\nolimits_1 = m\sum_{r=1}^{m} t^r_{i-n-j,i-n}t^r_{i-n,i} - \left(\sum_{r=1}^{m} t^r_{i-n-j,i-n}\right)\left(\sum_{r=1}^{m} t^r_{i-n,i}\right),$$

and

$$\sum\nolimits_2 = m\sum_{r=1}^{m}(t^r_{i-n-j,i-n})^2 - \left(\sum_{r=1}^{m} t^r_{i-n-j,i-n}\right)^2.$$

$$b_{t_{i,i-n,i-n-j}} = \frac{\sum_{r=1}^{m}\left(t^r_{i-n,i} - a_{t_{i,i-n,i-n-j}}t^r_{i-n-j,i-n}\right)}{m}. \quad (4)$$

$$t^{r\prime}_{i-n,i} = \left(\sum\nolimits_3\right) \Big/ \sum_{j=1}^{k} w_{t_{i,i-n,i-n-j}}, \quad (5)$$

where $\sum_3 = \sum_{j=1}^{k} w_{t_{i,i-n,i-n-j}} \times (a_{t_{i,i-n,i-n-j}}t^r_{i-n-j,i-n} + b_{t_{i,i-n,i-n-j}})$.

4) Decentralized Computing Devices: The decentralized computing devices contain multiple node devices, and each node device can train the multiple linear regression models. In training stage, the k weighted linear regression models are evenly distributed to each node device according to the map functions in MapReduce framework. The slope (e.g. $a_{t_{i,i-n,i-n-j}}$), intercept (e.g. $b_{t_{i,i-n,i-n-j}}$), and weight (e.g. $w_{t_{i,i-n,i-n-j}}$) of each multiple linear regression model can be calculated by the decentralized computing devices and would be stored in the cache database servers for further real-time analyses.

5) Combined Node Devices: The combined node devices can receive the data from the decentralized computing devices and integrate these data according to the reduce functions in the MapReduce framework for obtaining the analysis results. In this case, the combined node devices can receive the k weighted linear regression models and related parameters (i.e. slopes, intercepts, and weights) of these models from decentralized computing devices. The combined node devices can perform Equation (2) to calculate the predicted travel time from the $(i\text{–}n)$-th station to the i-th station.

6) Cache Database Servers: The cache database servers store the computing results and related parameters (i.e. the slopes, intercepts, and weights of k weighted linear regression models) from the decentralized computing devices for further analyses. These parameters can be adjusted in accordance with only newly-added records for accelerated computation.

III. The Proposed Method

The proposed multiple linear regression method based on stream homomorphic encryption computing includes four stages, i.e. (1) data preprocessing stage, (2) data access stage, (3) data computing stage, and (4) result processing stage.

A. Data Preprocessing Stage

In the data preprocessing stage, the two steps (1) retrieving data and (2) encrypting data are performed to collect data and store encrypted data for further analyses.

1) Retrieving Data: The data processing devices collect and retrieve the records from the web service servers and the database servers, for instance, OBU 1 (i.e. an user device) arrives at Station 1, Station 2, and Station 3 at 09:00:00, 09:03:20, and 09:07:00, respectively; OBU 2 (i.e. an user device) arrives at Station 1, Station 2, and Station 3 at 10:00:00, 10:04:00, and 10:08:10, respectively; OBU 3 (i.e. an user device) arrives at Station 1, Station 2, and Station 3 at 11:00:00, 11:03:30, and 11:07:20, respectively; OBU 4 (i.e. an user device) arrives at Station 1 and Station 2 at 12:00:00 and 12:03:40, respectively (as shown in TABLE I). When an OBU arrives at each station, it will report its location and time information to the web service servers and database servers through middleware (e.g. RESTful APIs). The data processing devices analyze these records and calculate the travel time between each pair of stations, for instance, the travel time $t_{1,2}$ from Station 1 to Station 2 taken by OBU1 is 200 seconds, and the travel time $t_{2,3}$ from Station 2 to Station 3 taken by OBU1 is 220 seconds (as shown in TABLE II).

TABLE I
ARRIVAL TIMES

	Station 1	Station 2	Station 3
OBU 1	09:00:00	09:03:20	09:07:00
OBU 2	10:00:00	10:04:00	10:08:10
OBU 3	11:00:00	11:03:30	11:07:20
OBU 4	12:00:00	12:03:40	09:07:00

TABLE II
TRAVEL TIME BETWEEN EACH PAIR OF STATIONS (UNIT: SECOND)

	From Station 1 to Station 2	From Station 2 to Station 3
OBU 1	200	220
OBU 2	240	250
OBU 3	210	230
OBU 4	220	

2) Encrypting Data: After calculating the travel time between each pair of stations by data processing devices, the related parameter values which include $t_{1,2} \times t_{2,3}$, $t_{1,2}$, $t_{2,3}$, and $t_{1,2} \times t_{1,2}$ are calculated by data processing devices (as shown in TABLE III). Furthermore, an encryption algorithm with a private key p, a public key q, and a random integer value z is performed to encrypt the related parameter values by Equation (6). In this case, the private key p is 39,916,801; the public key q is 112,909; the random integer value z is 7. For instance, the plaintext 44,000 is encrypted by Equation (6) as the ciphertext 279,461,607. The ciphertexts of the related parameter values in TABLE III are shown in TABLE IV.

$$f(x) \equiv (x + p \times z) \pmod{p \times q}. \tag{6}$$

TABLE III
RELATED PARAMETER VALUES

	$t_{1,2} \times t_{2,3}$	$t_{1,2}$	$t_{2,3}$	$t_{1,2} \times t_{1,2}$
OBU 1	44,000	200	220	40,000
OBU 2	60,000	240	250	57,600
OBU 3	48,300	210	230	44,100

TABLE IV
RELATED PARAMETER VALUES AFTER ENCRYPTION

	$t_{1,2} \times t_{2,3}$	$t_{1,2}$	$t_{2,3}$	$t_{1,2} \times t_{1,2}$
OBU 1	279,461,607	279,417,807	279,417,807	279,457,607
OBU 2	279,477,607	279,417,847	279,417,847	279,475,207
OBU 3	279,465,907	279,417,817	279,417,817	279,461,707

B. Data Access Stage

The data processing devices can store the ciphertexts of records to the decentralized database devices. In this case, the data processing devices can store the encrypted related parameter values (i.e. the values in TABLE IV) in the decentralized database devices to avoid the risks of data leakage when the records in these database devices are stolen.

C. Data Computing Stage

The data computing stage contains the two steps: (1) selecting data mining method and (2) decentralized computing are performed to calculate encrypted data.

1) Selecting Data Mining Method: The administrator devices can access the data mining devices via the data analysis devices and select the preferred data mining method. The multiple linear regression method is selected to analyze and build the multiple linear regression models in further steps.

2) Decentralized Computing: The decentralized computing devices can perform the selected data mining method (i.e. the multiple linear regression method) and directly process the ciphertext by homomorphic encryption. The summaries of related encrypted parameter values are calculated in TABLE V, and the parameters of multiple linear regression model (i.e. the slope a and the intercept b) can be estimated by these related encrypted parameter values.

TABLE V
THE SUMMARY OF RELATED PARAMETER VALUES AFTER ENCRYPTION

	$t_{1,2} \times t_{2,3}$	$t_{1,2}$	$t_{2,3}$	$t_{1,2} \times t_{1,2}$
Summary	838,405,121	838,253,471	838,253,471	838,394,521

D. Result Processing Stage

The result processing stage consists of the four steps: (1) storing summaries (2) decrypting data, (3) predicting travel time, and (4) showing results are performed to obtain results.

1) Storing Summaries: The data of OBU 1, OBU 2 and OBU 3 have been summed up, and the summaries of related encrypted parameter values can be temporarily stored in the cache database devices for accelerating computation.

2) Decrypting Data: The decentralized computing devices can send the summaries of related encrypted parameter values to the combined node devices, which can then perform data combination and decrypt the ciphertexts. A decryption algorithm with the same private key p, the same public key q, and the same random integer value z is performed to decrypt the ciphertexts by Equation (7). In this case, the private key p is 39,916,801; the public key q is 112,909; the random integer value z is 7. Then the ciphertext 838,405,121 is decrypted by Equation (7) as the plaintext 152,300. The plaintexts of the related summarized encrypted parameter values in TABLE V are shown in TABLE VI.

$$g(x) \equiv x \pmod{p}. \tag{7}$$

TABLE VI
THE SUMMARY OF RELATED PARAMETER VALUES AFTER DECRYPTION

	$t_{1,2} \times t_{2,3}$	$t_{1,2}$	$t_{2,3}$	$t_{1,2} \times t_{1,2}$
Summary	152,300	650	700	141,7001

3) Predicting Travel Time: The decrypted data and the number of data (i.e. $m = 3$) are adopted to calculate the parameters of multiple linear regression model (i.e. a and b) by Equations (8) and (9). The predicted travel time from Station 2 to Station 3 taken by OBU 4 is estimated about 236 seconds by Equation (10). Therefore, the arrival time of Station 3 is predicted to be 12:07:36.

$$a = \frac{3 \times 152300 - 650 \times 700}{3 \times 141700 - 650^2} = 0.730769. \tag{8}$$

$$b = \frac{1}{3}\left(700 - 650 \times a\right) = 75. \tag{9}$$

$$t = 220 \times a + 75 = 235.76918 \approx 236 \tag{10}$$

4) Showing Results: The combined node devices can transmit the prediction results to the data analysis devices, and the data analysis devices can notify the administrator devices for presenting the prediction results on the administrator devices. The administrator devices can receive the predicted results and show that the predicted arrival time of Station 3 is 12:07:36.

IV. EVALUATION AND ANALYSES

To evaluate the proposed method, the computation costs of RSA (Rivest-Shamir-Adleman) [1] and ECC (elliptic curves cryptography) [2] were compared in TABLE VII. In our experiments, the number of records is m. The computation costs of encryption and summation by these methods (i.e. the proposed method, RSA, and ECC) are $4 \times 4 \times m$ and $4 \times (m-1)$, respectively. There is a notable difference between the computation costs of decryption by the proposed method and other methods (i.e. RSA and ECC). The results shows that the computation cost of decryption by the proposed method was only one (i.e. time complexity $O(1)$), but the computation cost of decryption by RSA and ECC was $4 \times m$ (i.e. time complexity $O(m)$). Therefore, the performance of the proposed method is higher than other methods.

TABLE VII
COMPARISON OF COMPUTATION COSTS

Computation Operation	Proposed Method	RSA or ECC
Encryption (From TABLE III to TABLE IV)	$4 \times 4 \times m$	$4 \times 4 \times m$
Decryption Based on RSA or ECC	0	$4 \times m$
Summation (From Table 5 to Table 6)	$4 \times (m-1)$	$4 \times (m-1)$
Decryption Based on The Proposed Method	1	0
Total Computation Costs	$20 \times (m-3)$	$24 \times (m-4)$

V. CONCLUSIONS AND FUTURE WORK

Due to the privacy application development of the Internet of things, the needs of real-time and massive secure data calculations and analyses are an important issue. Therefore, this study proposes a multiple linear regression system coupled with relevant method based on stream homomorphic encryption computing. This proposed system includes the three subsystems: (1) terminal subsystem, (2) data access subsystem, and (3) data computing subsystem. This proposed method includes four stages: (1) data preprocessing stage, (2) data access stage, (3) data computing stage, and (4) result processing stage. The collected records can be encrypted and calculated with the cache mechanism to perform real-time streaming analyses, and the homogenous encryption method and multiple linear regression method are integrated to directly analyze ciphertexts in the database for the applications of traffic information prediction.

In the future, the homogenous encryption method can be applied to social networks [3] and be integrated with other data mining methods to enhance information security in the Internet of things environments.

ACKNOWLEDGMENT

This work was partially supported by the National Natural Science Foundation of China (Nos. 61906043, 61877010, 11501114, and 11901100), Fujian Natural Science Funds (No. 2019J01243), Funds of Education Department of Fujian Province (No. JAT190026), and Fuzhou University (Nos. 510872/GXRC-20016, 510930/XRC-20060, 510730/XRC-18075, 510809/GXRC-19037, 510649/XRC-18049, and 510650/XRC-18050). The authors would like to thank the chairs and valuable comments by reviewers of the IS3C 2020.

REFERENCES

[1] Y. Xu, S. Wu, M. Wang, Y. Zou, "Design and implementation of distributed RSA algorithm based on Hadoop," Journal of Ambient Intelligence and Humanized Computing, vol. 11, pp. 1047–1053, 2020.

[2] S. Sridhar, S. Smys, "Hybrid RSAECC based secure communication in mobile cloud environment," Wireless Personal Communications, vol. 111, pp. 429–442, 2020.

[3] H. Chen, H. Jin, S. Wu, "Minimizing Inter-server Communications by Exploiting Self-similarity in Online Social Networks," IEEE Transactions on Parallel and Distributed Systems, vol. 27, no. 4, pp. 1116–1130, 2016.