2023 International Conference on Consumer Electronics - Taiwan (ICCE-Taiwan) | 979-8-3503-2417-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCE-Taiwan58799.2023.10226978

Homomorphic Encryption-Based Similarity Calculation for Mobile Data Usage Prediction

Abel C. H. Chen*, Chia-Yu Lin†, and Yueh-Ting Lai*

* *Telecommunication Laboratories, Chunghwa Telecom Co., Ltd., Taoyuan, Taiwan*

† *Department of Computer Science and Information Engineering, National Central University, Taoyuan, Taiwan*

Corresponding Author: Chia-Yu Lin (sallylin0121@ncu.edu.tw)

***Abstract*—For improving security, this study proposes a mobile data usage prediction method based on homomorphic encryption (HE)-based similarity calculation without plaintext storage. The cosine similarity, angular similarity, Tanimoto similarity, and Euclidean similarity based on HE according to ElGamal cryptography are proposed and combined in a machine learning method to consider the mobile data usage during the last periods for predicting the mobile data usage during the next period.**

***Index Terms*—Homomorphic encryption, similarity, mobile data usage prediction**

I. Introduction

Due to increasingly serious network attacks, systems and methods' security and privacy issues have been emphasized. Some homomorphic encryption (HE) methods have been proposed to provide the addition and multiplication results of ciphertexts. The decrypted results based on HE methods can be the same as the addition and multiplication results of plaintexts [1]. Thus, the ciphertexts can be stored in the cloud environments for analysis to improve security.

Although some mobile data usage prediction (MDUP) methods [2] have been proposed, the plaintexts of mobile data are used for analysis with potential security issues. Therefore, this study enhances the authors' study [2] to propose a machine learning method with HE-based similarity calculation based on ElGamal cryptography for MDUP.

II. Related Work

A. ElGamal Cryptography

In the ElGamal encryption system, a public key Q is generated by a private key q based on a prime numer p by Eq. (1). The plaintext $x_{i,k}$ can be encrypted by the encryption function $E\left(x_{i,k}\right)$ to obtain the ciphertext $\left(c_1, c_{i,k}\right)$ by Eq. (2) and (3) according to a generator g and a random number r. Furthermore, the ciphertext $\left(c_1, c_{i,k}\right)$ can be decrypted based on the private key q by Eq. (4) [3].

$$Q = g^q \mod p \tag{1}$$

$$c_1 = g^r \mod p \tag{2}$$

$$E\left(x_{i,k}\right) = c_{i,k} = x_{i,k}R, \text{ where } R = Q^r \mod p \tag{3}$$

$$D\left(c_{i,k}\right) = c_{x_{i,k}}\left(c_1^q\right)^{-1} \mod p = x_{i,k} \tag{4}$$

B. Similarity Calculation

This subsection shows the principles of cosine similarity, angular similarity, and Tanimoto similarity.

1) Cosine Similarity: The cosine similarity of matrices X_i and X_j can be measured by Eq. (5) based on the multiplication of X_i and X_j and the inner product of each matrix [4].

$$S_c\left(X_i, X_j\right) = \frac{X_i X_j}{\|X_i\|\|X_j\|} \tag{5}$$

2) Angular Similarity: The angular similarity of matrices X_i and X_j can be measured by Eq. (6) based on the arccosine of cosine similarity and the circular constant π [4].

$$S_a\left(X_i, X_j\right) = 1 - \frac{2 \times \arccos S_c\left(X_i, X_j\right)}{\pi} \tag{6}$$

3) Tanimoto Similarity: The Tanimoto similarity of matrices X_i and X_j can be measured by Eq. (7) based on the union and intersection of matrices X_i and X_j. Furthermore, the Tanimoto distance can be derived by Eq. (8) that includes $\|X_i - X_j\|^2$ (i.e. likes the Euclidean distance) [4].

$$S_t\left(X_i, X_j\right) = \frac{X_i X_j}{\|X_i\|^2 + \|X_j\|^2 - X_i X_j} \tag{7}$$

$$\begin{aligned} 1 - S_t\left(X_i, X_j\right) &= 1 - \frac{X_i X_j}{\|X_i\|^2 + \|X_j\|^2 - X_i X_j} \\ &= \frac{\|X_i - X_j\|^2}{\|X_i\|^2 + \|X_j\|^2 - X_i X_j} \end{aligned} \tag{8}$$

III. The Proposed System and Method

A. The Proposed System

The proposed method can collect the mobile data usage amount of each user through the core network in cellular networks [2]. Furthermore, this study can encrypt the collected mobile data usage amount and store the encrypted information into the database server.

B. The Proposed Method

The MDUP method is designed based on k-nearest neighbors (kNN) [2]. Therefore, similarities and distances between each two matrices are important factors in kNN. For obtaining HE-based similarity calculation, this study considers ElGamal cryptography to combine the multiplicative HE with cosine similarity, angular similarity, Tanimoto similarity, and

Euclidean similarity. Each n-dimension plaintext is detnoed as X_i, and the ciphertext of X_i is encrypted as C_i by (3).

1) Homomorphic Encryption-Based Cosine Similarity (HE-CS): HE-CS based on Eq. (9) is proposed to decrypt the summaries of ciphertexts (i.e. matrices C_i and C_j) for obtaining the cosine similarity of X_i and X_j [4].

$$\begin{aligned}\frac{D\left(C_iC_j\right)}{D\left(\|C_i\|\|C_j\|\right)} &= \frac{\sum\limits_{k=1}^{n}\frac{c_{i,k}}{c_1^q}\frac{c_{j,k}}{c_1^q} \mod p}{\sqrt{\sum\limits_{k=1}^{n}\left(\frac{c_{i,k}}{c_1^q}\right)^2}\sqrt{\sum\limits_{k=1}^{n}\left(\frac{c_{j,k}}{c_1^q}\right)^2} \mod p}\\ &= \frac{\sum\limits_{k=1}^{n}x_{i,k}x_{j,k}}{\sqrt{\sum\limits_{k=1}^{n}x_{i,k}^2}\sqrt{\sum\limits_{k=1}^{n}x_{i,k}^2}} = S_c\left(X_i, X_j\right)\end{aligned} \tag{9}$$

2) Homomorphic Encryption-Based Angular Similarity (HE-AS): HE-AS based on Eq. (10) is proposed based on HE-CS to measure the arccosine of cosine similarity for obtaining the angular similarity of X_i and X_j [4].

$$\begin{aligned}1 - \frac{2 \times \arccos\frac{D(C_iC_j)}{D(\|C_i\|\|C_j\|)}}{\pi} &= 1 - \frac{2 \times \arccos S_c\left(X_i, X_j\right)}{\pi}\\ &= S_a\left(X_i, X_j\right)\end{aligned} \tag{10}$$

3) Homomorphic Encryption-Based Tanimoto Similarity (HE-TS): HE-TS based on Eq. (11) is proposed to decrypt the summaries of ciphertexts (i.e. matrices C_i and C_j) for obtaining the Tanimoto similarity of X_i and X_j [4].

$$\begin{aligned}&\frac{D\left(C_iC_j\right)}{D\left(\|C_i\|^2\right) + D\left(\|C_j\|^2\right) - D\left(C_iC_j\right)}\\ &= \frac{X_iX_j}{\|X_i\|^2 + \|X_j\|^2 - X_iX_j} = S_t\left(X_i, X_j\right)\end{aligned} \tag{11}$$

4) Homomorphic Encryption-Based Euclidean Similarity (HE-ES): The Euclidean distance is defined in Eq. (12), and the Euclidean similarity can be estimated by Eq. (13). Therefore, HE-ES based on Eq. (14) is proposed to decrypt the squared difference of ciphertexts (i.e. matrices C_i and C_j) for obtaining the Euclidean similarity of X_i and X_j [4].

$$\|X_i - X_j\|^2 \tag{12}$$

$$S_e\left(X_i, X_j\right) = \frac{1}{1 + \|X_i - X_j\|^2} \tag{13}$$

$$\begin{aligned}\frac{1}{1 + \|C_i - C_j\|^2} &= \frac{1}{1 + D\left(\|C_i - C_j\|^2\right)}\\ &== \frac{1}{1 + \|X_i - X_j\|^2} = S_e\left(X_i, X_j\right)\end{aligned} \tag{14}$$

IV. Experiments and Comparison

In experiments, the mobile data usage records were collected from 6,722 users of Chunghwa Telecom mPro 650 Project for the evaluation of the proposed methods. The length of a period is five days; the mobile data usage amount of each user during the last five periods is selected as the inputs of the machine learning method, and the mobile data usage amount of each user during the next period is adopted as the output for prediction evaluation.

The encryption time based on Eq. (3) for 6,722 users and six periods is 3.72s; the decryption time of HE-CS, HE-AS, HE-TS, and HE-ES are 7805.55s, 5407.57s, 417.36s, and 2360.86s. The evaluation factor of mean absolute percentage error (MAPE) is selected for evaluation, and the MAPEs of the proposed methods based on kNN under different numbers of k are shown in Table I. The proposed HE-TS has provided higher performance for secured mobile data usage prediction.

TABLE I
The MAPEs of the Proposed Methods

Method	k=1	k=2	k=3	k=4	k=5
HE-CS	26.39%	29.39%	29.73%	30.28%	30.57%
HE-AS	26.22%	29.25%	29.62%	30.18%	30.48%
HE-TS	8.62%	9.86%	10.44%	10.88%	11.20%
HE-ES [2]	31.81%	37.09%	33.84%	32.41%	33.45%

V. Conclusions

This study proposes a prediction method based on HE-based similarity according to multiplicative HE. The proposed methods have been proved based on mathematical models, and the practical MDUP is evaluated in accordance with the data from Chunghwa Telecom. In the future, the proposed HE-based similarity can be combined with other machine learning methods for improving the accuracy of prediction.

Acknowledgements

This work is sponsored by the National Science and Technology Council (NSTC) under the project NSTC 111-2622-8-A49-023 and NSTC 110-2222-E-008-008-MY3.

References

[1] S. Jumonji, K. Sakai, M. -T. Sun and W. -S. Ku, “Privacy-Preserving Collaborative Filtering Using Fully Homomorphic Encryption,” IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 3, pp. 2961–2974, 2023.

[2] Y. -T. Lai, Y. -P. Wu, C. -H. Yu, F. -S. Lu and C. -H. Chen, “Mobile Data Usage Prediction System and Method,” 2017 31st International Conference on Advanced Information Networking and Applications Workshops (WAINA), Taipei, Taiwan, 2017, pp. 484–486.

[3] T. Elgamal, “A Public Key Cryptosystem and a Signature Scheme Based on Discrete Logarithms,” IEEE Transactions on Information Theory, vol. 31, no. 4, pp. 469–472, 1985.

[4] Abel C. H. Chen, “Similarity Calculation Based on Homomorphic Encryption,” arXiv, ARXIV.2302.07572, 2023, doi: 10.48550/arxiv.2302.07572.