Elliptic Curve Cryptography (ECC) Implementation on Naïve Bayes Classifier for Privacy-Preserving Data Mining

Choirunnisa Fatima 13512084
Program Studi Teknik Informatika
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia
choirunnisa.fatima@gmail.com

***Abstract*—As now people consider the importance of preserving their privacy data, this leads to some restriction in data mining. Data mining is used to extract some knowledge from large amount of database. However, these data are distributed among different locations and often involve other parties to mine them. Therefore it could violate privacy and security issues if company reveal data or information to other parties. Recently researches of privacy preserving data mining has been conducted to address privacy problem in data mining. Cryptographic techniques are one of the solution. In this paper we propose an implementation of Elliptic Curve Cryptography (ECC) on Naïve Bayes classifier using 1-out-of-n Oblivious Transfer protocol for Privacy-Preserving Data Mining.**

***Index Terms*—Elliptic Curve Cryptography, Naïve Bayes, Oblivious Transfer, Privacy Preserving Data Mining.**

I. INTRODUCTION

Data mining has been operated by gathering all data into a central site, then running algorithm to the data. The data are often gathered from multi parties that privacy is considered can prevent this approach. For example, government wish to conduct a medical research. Therefore they have to gather and collect medical data from different hospitals. However, they must protect the data from violating any privacy issues, they must not reveal the data of one hospital to other hospital. Yet they also must not reveal the data for their selves.

Privacy preserving data mining (abbreviated as PPDM) deals with this issue, that is protecting the privacy of data or sensitive knowledge without sacrificing the purpose of data mining. People have become aware of the privacy intrusions on their personal data and are afraid to share their personal information. However, organizations or companies wish to do data mining on their database, so that they can improve service performance or make a good product. With joint data mining, they can even get better information or knowledge from joining their databases. The problem is how they could do data mining on their joining databases, such as build decision tree, data clustering, identify association rules, without learning other databases and only learning the output of data mining process. Cryptography is a tool to overcome this problem. This problem of privacy-preserving data mining is actually a special case in cryptography called Secure Multi-party Computation.

This problem deals with a number of parties and their private inputs who wishes to jointly compute some function on their inputs. This computation should have property that parties only learn the output of computation and nothing else. Loosely speaking, we need a protocol that guarantees this property to solve privacy-preserving data mining problems.

A simple secure computation protocol which guarantees this property is oblivious transfer. This protocol is involving only two parties. It is a basic building block of many cryptographic protocols for secure computation. The most famous variant of oblivious transfer is 1-out-of-2 oblivious transfer, which was suggested by Eve, Goldreich and Lempel [9] and another variant of oblivious transfer that was suggested by Rabin [8]. However, in this paper we use another variant of oblivious transfer, that is 1-out-of-n oblivious transfer. We will use protocol for 1-out-of-n oblivious transfer collaborated with elliptic curve cryptography which has been proposed by Parakh [1]. Further explanation of Parakh's protocol is discussed in section III.

II. ELLIPTIC CURVE CRYPTOGRAPHY

Elliptical Curve Cryptography (abbreviated as ECC) is a public key cryptography.

The advantage of ECC compared to RSA is that it offers equal security for a smaller bit size, thereby reducing processing overhead. ECC can only encrypt and decrypt a point on the curve and not messages, therefore message should be encoded into point. Padma's paper [2] discusses Koblitz's method to represent a message to a point and vice-versa.

The general form of the elliptic curve equation used is $y^2 \ mod \ p = (x^3 + ax + b) \ mod \ p$ where $4a^3 + 27b^2 \ mod \ p \neq 0$. That is the equation of elliptic curve on a prime field F_p. The elements of the finite field are integers between 0 and $p - 1$. All operations on the finite field are modular arithmetic, thus involves only integers between 0 and $p - 1$. These operations of points, however, have specific rules.

1) Point Addition
   Let J and K are two points on elliptic curve over F_p such that $J = (x_J, y_J)$ and $K = (x_K, y_K)$. Let $L = J + K$ where $L = (x_L, y_L)$, then
   $x_L = s^2 - x_J - x_J \; mod \; p$
   $y_L = -y_J + s(x_J - x_L) \; mod \; p$
   s is the slope of the line through J and K, $s = \frac{y_J - y_K}{x_J - x_K} mod \; p$.
   If $K = -J$ then $J + K = O$, where O is the point at infinity. If $K = J$ then $J + K = 2J$ then point doubling operations are used.
2) Point Subtraction
   Let J and K are two points on elliptic curve over F_p such that $J = (x_J, y_J)$ and $K = (x_K, y_K)$. Then $J - K = J + (-K)$ where $-K = (x_K, -y_K \; mod \; p)$.
3) Point Doubling
   Let J is a point on elliptic curve over F_p such that $J = (x_J, y_J)$. Let $L = 2J$ where $L = (x_L, y_L)$, then
   $x_L = s^2 - 2x_J \; mod \; p$
   $y_L = -y_J + s(x_J - x_L) \; mod \; p$
   s is the tangent at point J and a is a parameter of elliptic curve equation, $s = \frac{3{x_J}^2 + a}{2y_J} mod \; p$.
   If $y_J = 0$ then $2J = O$, where O is point at infinity.

The security of ECC depends on the difficulty of Elliptic Curve Discrete Logarithm Problem. The difficulty is to obtain k scalar from $kP = Q$, given P and Q are two points on an elliptic curve. k is said to be the discrete logarithm of Q to the base P. Therefore the main operation involved in ECC is point multiplication.

III. 1-OUT-OF-N OBLIVIOUS TRANSFER

Oblivious transfer protocol is a cryptography protocol in which a sender sends a message to a receiver, but remains oblivious as to which piece of message has been sent. For example, Alice has two secrets A_0 and A_1, and Bob has a bit b. Bob wants to receive A_b without Alice knowing b. Alice has to ensure that Bob receives only A_b and not A_{1-b}. As Bob only receives 1 from 2 secrets, this protocol had been named as 1-out-of-2 oblivious transfer protocol.

A 1-out-of-n oblivious transfer protocol can be defined as a natural generalization of a 1-out-of-2 oblivious transfer protocol. Specifically, Alice has n secrets, Bob has index i and Bob wants to receive i-th among Alice's secrets without Alice knowing i. Parakh [1] had proposed protocol for communication efficient 1-out-of-n oblivious transfer using elliptic curve cryptography. The protocol not only reduces burden on communication over network but also reduces computational burdens for both sender and receiver. The security of the protocol is predicated on the difficulty of elliptic curve discrete log problem.

In Parakh's [1] proposed protocol, Alice and Bob made agreements upon an elliptic curve to use and point G. Alice randomly and uniformly chooses k_0 and k_1 and generates two points P_0 and P_1 such that $P_0 = k_0 G$ and $P_1 = k_1 G$. These values must not be known to Bob. Inputs for the protocol are Alice's secrets S_0, S_1,…, S_{n-1} and Bob's choice of index $i \; \epsilon \; \{0,1, \dots, n-1\}$. Then the protocol steps are proceed as follows.

1) Alice chooses a point $C = P$ on elliptic curve.
2) Alice chooses randomly and uniformly a number r from the field and generates rG.
3) Alice sends C and rG to Bob.
4) Bob chooses randomly and uniformly a number k and sets $PK_i = kG$.
5) Bob computes the decryption key $krG = rPK_i$.
6) If $i \neq 0$, Bob computes $PK_0 = iC - PK_i$.
7) Bob sends PK_0 to Alice.
8) Alice computes rPK_0 and for all $1 \leq j \leq n-1$, she computes $rPK_j = rjC - rPK_0$.
9) Alice uses rPK_j as encryption key to encrypt S_j: $E(S_j)$ and sends them to Bob.
10) Bob chooses $E(S_i)$ and decrypts it using rPK_i to extract S_i.

Note that when $i = j$ then $rPK_j = rjC - rPK_0 = riC - riC - rPK_i = rPK_i$. Thus, Bob will be able to decrypt $E(S_i)$ and get secret S_i. Otherwise, Bob won't get any information.

IV. NAÏVE BAYES CLASSIFIER

The Naïve Bayes classifier is a simple probabilistic classifier which is based on Bayes theorem with strong and naïve independence assumptions. The Naïve Bayes classifier is a popular method for text categorization, that is the problem to judge documents are belonging to one category or the other. All Naïve Bayes classifiers assume that the value of certain feature is independent to the other feature. An advantage of Naïve Bayes is that it requires only a small amount of training data to estimate the parameters necessary for classification.

Naïve Bayes is a conditional probability model. Given a problem dataset to be classified, we must represent the problem into probability model. The conditional probability model is something like $p(C_k|x_1, \dots, x_n)$ for each of k possible outcomes or classes, where $\boldsymbol{x} = (x_1, \dots, x_n)$ is a vector representing n features assigned to the probability model. Using Bayes' theorem, we can decompose the conditional probability as $p(C_k|\boldsymbol{x}) = \frac{p(C_k)p(\boldsymbol{x}|C_k)}{p(\boldsymbol{x})}$. As $p(\boldsymbol{x})$ is identical for all classes, therefore it can be ignored. This yields a discriminant function for each k class, $f_k(\boldsymbol{x}) = p(C_k)p(\boldsymbol{x}|C_k)$. Thus, the Bayes classifier is the function that assigns a class label $\hat{y} = C_k$ for some k as follows:

$$\hat{y} = \underset{k \in \{1, \dots, K\}}{argmax} p(C_k) \prod_{i=1}^{n} p(x_i|C_k).$$

The classifier finds the maximum a posteriori probability (MAP) hypothesis given example x. However, estimation of $p(\boldsymbol{x}|C_k)$ is hard when the feature space is high-dimensional. Therefore, approximations are often used.

V. THE PROPOSED ALGORITHM

We consider a scenario where a party wish to classify an instance of their private database using data train from private database of other party. Since the databases are confidential, neither party is willing to reveal any of the information to the other. We will show how data mining problem of naïve bayes can be easily computed, without no party learning anything other than the output. We modified Parakh's protocol for 1-out-of-n oblivious transfer to build our proposed algorithm.

A. Assumptions

Let $\boldsymbol{X} = (X_1, \dots, X_l)$ be a vector of observed variables, called features, where each feature takes value from domain D_i. Capital letters, such as X_i will denote variables, while lower-case letters, such as x_i will denote their values.

Alice has a dataset containing $x_{i,j}$ where $1 \leq i \leq l$ and $1 \leq j \leq m$. i denotes index of feature and j denotes index of instance of dataset. Bob has an instance containing x_k where $1 \leq k \leq l$, k denotes index of feature.

Alice doesn't know which variable of Bob's instance has missing value (class) and she doesn't know any other value of Bob's instance. Bob doesn't know any value of Alice's dataset. Bob will learn the conditional probability from Alice's dataset according his instance. They made agreement on features and domain parameters of elliptic curve. No other party involved in this protocol.

The domain parameters of elliptic curve are **p**, **a**, **b**, **G**, **n**, and curve $y^2 \ mod \ p = (x^3 + ax + b) \ mod \ p$. G is the generator point, n is the order of the elliptic curve. We use a cryptographic hash function, such as SHA-1, to convert string into integer.

B. Algorithm

The steps for our proposed algorithm proceeds as follows.

1) Alice chooses a point $C = P$ on the elliptic curve.
2) Alice chooses a private key a, a randomly selected integer less than n. Then generates public key $A = aG$.
3) Bob chooses a private key b, a randomly selected integer less than n. Then generates public key $B = bG$.
4) Alice sends C and A to Bob.
5) Bob computes the decryption key $bA = abG$.
6) Bob chooses an index c which is the index of feature to be the class. Clearly $1 \leq c \leq l$. Then computes $PK = \ cC - B$. Bob sends PK to Alice.
7) Alice computes aPK and for all $1 \leq i \leq l$, she computes $aPK_i = aiC - aPK$. Alice uses aPK_i as encryption key to encrypt her dataset $x_{i,j}$. Then sends them all to Bob.
8) When Bob receives the encrypted dataset, he chooses $x_{i=c,j}$ and decrypts it using bA. Therefore he will only get information on the feature he chose to be class.
9) Bob has an instance in form of vector $\{x_k | 1 \leq k \leq l\}$. He calculates $e_k = HASH(x_k)$, where e is an integer result of hash function.
10) Alice also calculates $e_{i,j} = HASH(x_{i,j})$ for all elements in her dataset.
11) Bob computes $PK_k = \ e_k C - B$. Then sends vector $\{PK_k | 1 \leq k \leq l\}$ to Alice.
12) Alice computes aPK_i from Bob's vector and for all $e_{i,j}$ in her dataset she computes $aPK_{e_{i,j}} = ae_{i,j}C - aPK_i$, where $1 \leq i \leq l$ and $1 \leq j \leq m$. Alice uses $aPK_{e_{i,j}}$ as encryption key to encrypt her dataset $x_{i,j}$. Then sends them all to Bob.
13) Again Bob receives an encrypted dataset from Alice. Then he decrypts it using key bA.
14) At last, with two information he got, Bob could classify his instance using Naïve Bayes classifier

$$\hat{y} = \underset{k \in \{1, \dots, K\}}{argmax} p(C_k) \prod_{i=1}^{n} p(x_{i,j} = x_i | C_k).$$

C. Example

Alice and Bob made agreement on domain parameters of elliptic curve $y^2 \ mod \ p = (x^3 + ax + b) \ mod \ p$, those are:

a = -3
b = 5ac635d8aa3a93e7b3ebbd55769886bc651d06b0cc 53b0f63bce3c3e27d2604b
p = 11579208921035624876269744694940757353008 614341529031419553363130886709785395 1
G = 6b17d1f2e12c4247f8bce6e563a440f277037d812d eb33a0f4a13945d898c296, 4fe342e2fe1a7f9b8ee7eb4a7c0f9e162bce33576 b315ececbb6406837bf51f5
n = 15792089210356248762697446949407573529996955224135760342422259061068512044369

Suppose Alice has a dataset of playing tennis below.

outlook	temp	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

Table 1 Alice's Dataset

Bob has an instance with missing value below.

outlook	temp	humidity	windy	play
sunny	cool	high	true	?

Table 2 Bob's Instance to be Classified

After doing all first five steps, at 6th step, Bob chooses index $c = 5$. Then at 8th step Bob receives an encrypted dataset then decrypts it. The following table is dataset which Bob has decrypted.

outlook	temp	humidity	windy	play
				no
				no

				yes
				yes
				yes
				no
				yes
				no
				yes
				yes
				yes
				yes
				yes
				no

Table 3 Bob has decrypted the dataset

At 13th step Bob decrypts Alice's encrypted dataset. Then he joins the two datasets he received. Table below is Bob's final dataset.

outlook	temp	humidity	windy	play
sunny		high		no
sunny		high	true	no
		high		yes
		high		yes
	cool			yes
	cool		true	no
	cool		true	yes
sunny		high		no
sunny	cool			yes
				yes
sunny			true	yes
		high	true	yes
				yes
		high	true	no

Table 4 Bob's Final Dataset

With the dataset he has, Bob could classify his instance.

$$p(yes)p(sunny|yes)p(cool|yes)p(high|yes)p(true|yes)$$
$$= \frac{9}{14} \times \frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} = 0.0053$$
$$p(no)p(sunny|no)p(cool|no)p(high|no)p(true|no)$$
$$= \frac{5}{14} \times \frac{3}{5} \times \frac{1}{5} \times \frac{4}{5} \times \frac{3}{5} = 0.0206$$
$$\hat{y} = \underset{C_k \in \{yes, no\}}{argmax} p(C_k)p(sunny|C_k)p(cool|C_k) \; p(high|C_k)p(true|C_k)y$$
$$= no$$

Thus, the class of Bob's instance is no.

VI. Conclusion

We have presented the implementation of Elliptic Curve Cryptography on Naïve Bayes Classifier for Privacy-Preserving Data Mining using 1-out-of-n oblivious transfer. Due to properties of ECC, the security of this implementation is predicated on the difficulty of elliptic curve discrete log problem. However, this implementation is not efficient enough for high dimensional feature domains.

Future work can extend this implementation into more efficient algorithm which Alice and Bob only send and receive one message. Also modify this implementation so that could be used in another variant of naïve bayes classifier.

References

[1] A. Parakh. Communication Efficient Oblivious Transfer Using Elliptic Curves. In *IEEE 14th International Symposium on High-Assurance Systems Engineering*. 2012.

[2] Padma Bh, D. Chandravathi, P. Prapoorna Roja. Encoding and Decoding of a Message in the Implementation of Elliptic Curve Cryptography using Koblitz's Method. *(IJSE) International Journal on Computer Science and Engineering Vol. 02.* No. 05. 2010.

[3] Anoop MS. Elliptic Curve Cryptography An Implementation Guide. Available at http://informatika.stei.itb.ac.id/~rinaldi.munir/Kriptografi/2014-2015/ECC_Tut_v1_0.pdf.

[4] I. Rish. An Empirical Study of the Naïve Bayes Classifier. *IBM Research Report*. Nov 2, 2001.

[5] Y. Lindell, B. Pinkas. Secure Multiparty Computation for Privacy-Preserving Data Mining. In *The Journal of Privacy and Confidentiality (2009)*. 1, Number 1, pp. 59-98.

[6] A. Parakh. Oblivious Transfer Using Elliptic Curves. *Cryptologia*, 31:125-132, 2007.

[7] C. Cilton, et al. Tools for Privacy Preserving Distributed Data Mining. Available at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.10.417&rep=rep1&type=pdf

[8] M. O. Rabin. How to Exchange Secrets by Oblivious Transfer. Technical Memo TR-81, Aiken Computation Laboratory, 1981.

[9] S. Even, O. Goldreich and A. Lempel. A Randomized Protocol for Signing Contracts. *Communications of the ACM*, 28(6):637-647, 1985.

[10] S. Russell, P. Norvig. *Artificial Intelligence: A Modern Approach* (3rd ed.). Prentice Hall. 2003.

Pernyataan

Dengan ini saya menyatakan bahwa makalah yang saya tulis ini adalah tulisan saya sendiri, bukan saduran, atau terjemahan dari makalah orang lain, dan bukan plagiasi.

Bandung, 11 Mei 2015

Choirunnisa Fatima
13512084