International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 3 | ISSN : 2456-3307

Maintaining Security In Distributed Association Rule Mining Process Using Elliptic Curve Cryptography And Key Exchange Concept

J. Sumithra Devi[1], M. Ramakrishnan[2]
[1]Research Scholar, Department of Computer Science, Bharathiar University, Coimbatore, Tamil Nadu, India
[2]Chairperson, School of Information Technology, Madurai Kamaraj University, Madurai, Tamil Nadu, India

ABSTRACT

Association rule mining is one of the most important data mining processes which are used to generate useful patterns from huge volumes of data. It may so happen that the data source may not be from one sender and it is distributed across the world. In this distributed architecture, maintaining privacy while transmitting the data from originator to the data mining analyzer is a biggest issue as the data is highly confidential. Usage of cryptography is having a good impact on maintaining privacy in association rule mining process. Hence in this paper, elliptic curve cryptography based association rule mining process is presented. Experimental result shows that the difference of accuracy while using actual data and perturbed data is very less.
Keywords : Privacy Preservation,Security, Elliptic Curve Cryptography

I. INTRODUCTION

Enormous amount of data is being generated called as big data and this is purely because of rapid advancements in technology, computer applications and emergence of mobile/online social networks. Daily around 2.5 trillion bytes of data is produced every day and for the past two years, 85% of them have been produced[1]. Big data need more attention from the people, academic, business, government as they are having high potential to predict future trends, relationship, generate useful patterns and help us to take concrete future decisions[2].

To mine it to find important knowledge useful to support decision-making that could have a good impact on economic growth and technical innovations, we need the help of good decision making techniques, methods and tools. Data mining derives its name from the similarities between searching for valuable business information in a large database like finding linked products in gigabytes of store scanner data, mining a mountain for a vein of valuable ores[3]. Both processes require either shifting through an immense amount of material, or intelligently probing it to find exactly where the value resides[4].

Data mining techniques can yield the benefits of automation on existing software and hardware platforms, and can be implemented on new systems as existing platforms are upgraded and new products developed[5]. High speed makes it practical for users to analyze huge quantities of data. Larger databases, in turn, yield improved predictions. Faster processing means that users can automatically experiment with more models to understand complex data. When data mining tools are implemented on high performance parallel processing systems, they can analyze massive databases in minutes[6].

Recent advances in data mining technologies have increased the disclosure risks of sensitive data. Hence, the security issue has become, recently, a much more important area of research[7]. The problem in mining such confidential data is how the mining process can be done while maintaining the confidentiality of data (privacy-preservation), especially if the data is owned by several parties or agencies. This problem has raised such a new branch of data mining, called the Privacy-Preserving Data Mining (PPDM). In order to achieve the above goal, Elliptic Curve Cryptography (ECC) technique for a new efficient PPARM is proposed in this paper. Hence this paper is structured as follows: section 1 provides introduction to big data and association rule mining process. Section 2 presents a lucid literature survey used to get the basic ideas to precede our research. Basic concepts of elliptic curve cryptography are well discussed in section 3. Our proposed method was introduced in section 4 along with concepts of secure transmission and implementation. Experimental results and further discussion are presented in section 5. Paper ends by briefing the findings as conclusion in section 6.

II. RELATED WORK

Privacy preserving association rule mining is to find frequent itemsets in case of imprecise access to the original dataset and provide the association rules meeting the given support and confidence. The most famous algorithm is Mining Associations with Secrecy Constraints (MASK) and this algorithm is to map the original dataset into two-dimensional Boolean matrix and transforms the data with the Bernoulli probability model[8]. Data miners can get the transformed Boolean matrix and estimate the original support by the reconstruction algorithm to thereby generating frequent itemsets. It achieves privacy through the method of data interference and contains disadvantages. Privacy protection is not very effective as the transformed data and the original data are relevant and the value of the random parameter is subject to certain restrictions[9]. Improved the algorithm of Mining Association Security Konstraint(MASK) uses Randomized Disturbance and Reconstruction of Distributions to fulfill privacy requirement but the algorithm lacks in achieving time efficiency[10]. Single algorithm usage reduces the rate that suggested in hybrid approach. DPQR combines the techniques of data perturbation and query restriction and this method is better that MASK but it's implementation is tough as data to be analyzed should be in Boolean format and other data types are not allowed[11].

Randomized Response with Partial Hiding (RRPH) algorithm uses randomized parameters to interfere with data and exhibits better properties and efficiency compared to MASK[12]. RRPH is not well protected as first random parameter has not been distributed. Variations of this algorithm includes Privacy Association Rules Mining-Related Technology, multi-parameters randomized disturb algorithm, Partial Hiding Transition Probability Matrix [13].

PPARM using Hadoop framework was proposed by Jung et al. [12] which adds dummy as noise to the original transaction data. This technique is enough to prevent security violation but it slightly reduces the performance of the mining process[14]. This drawback can be eradicated in Hybrid Partial Hiding algorithm (HPH) which interfere with the original data, gives the frequent itemset generation algorithm and better provides the privacy protection in association rule mining.

Iqbal et al. [15] proposed PPDM for application related to XML documents. The proposed model is based on the Bayesian Network (BN). It finds sensitive rules and count the number which can be called reliable. BN approach which uses K2 algorithm and supports the Apriori algorithm for the AR can answer both previous questions. Lai et al [16]

proposed privacy preservation for cloud data owners for outsourcing the data process. The objective is to protect sensitive data and mining results. For this problem, they provided the first semantically secure solution with categorical data.

III. ELLIPTIC CURVE CRYPTOGRAPHY

Elliptic Curve Cryptography is an effective information hiding Public Key Cryptography that can reduces the use of bits with large numbers[17]. It is having widespread usage in so many security preserving applications and it was introduced by Neal Koblitz and Victor Miller in 1985. The Elliptic Curve systems use points on a curve called elliptic curves to generate key bits required for encryption and decryption processes[18].

So many variations of elliptic curves are recommended by international standard organizations or community such as ANSI that recommends Prime, C2pnb, and C2tnb elliptic curves; Standard for Efficient Cryptography Group (SECG) that recommends Secp elliptic curves; Brainpool group recommends Brainpool elliptic curves; and CDC Group of Darmstadt University, Germany that recommends PrimeCurve elliptic curves[19]. Curves in the real number, in the prime finite field GF(P), and curves in the binary finite field GF(2m). It is perceived that the ECC is able to meet the security standards with much smaller key size than that of the other systems.

IV. PROPOSED METHOD

Our proposed method works in three phases. Identification of different data sources and encryption of data using ECC before it is transferred on to the Data Warehouse is done at the first phase. During the next phase, data is decrypted and sent from different sources to be ready for transformation. Data is transformed to the format suitable for data warehouse. Data is loaded into the warehouse in the third phase and it uses data distortion to ensure privacy of sensitive information.

4.1 Secure Transmission

The performance of elliptic curve cryptography can be enhanced using finite fields of particular interest and it is represented by mod p (where p is a prime number). It chooses two non-negative integer values u and v such that u,v < p that satisfies the following condition

$$4u^3 + 27v^2 (\text{mod } p) \neq 0$$

The elliptic curve group, represented by p, whose elements are x ,y (both are non-negative integers and less than p) must satisfy the below condition

$$y^2 \equiv x^3 + ax + b(\text{mod } p)$$

The elliptic curve discrete logarithmic problem can be stated as follows

$$Q = xP$$

Where xP represents the point P on elliptic curve added to itself x number of times. From this we can easily calculate Q given x and P values.

4.2 Implementation

Let D1 and D2 be two distributed data sets. The datasets D1 and D2 are from two different senders S1 and S2 respectively. Let R1 be the receiver having the data warehouse. The first work is to encode the dataset records as m to be sent in point Pm. The point Pm is encrypted and decrypted. For any key exchange cryptosystem, a point G is required and an elliptic curve group is passed as parameters. The encryption cannot is simply performed on point Pm as not all the coordinates reside in $E_p(u, v)$.

Randomly sender S1 and S2 selects a private keys such that

$$P_A = n_A XG \quad || \quad P_B = n_B XG$$

Where P_A and P_B are private keys selected by first and second senders. To encrypt and send a message Pm to R1, a random positive integer x is used by the sender and generates a cipher text Cm such that

$$C_m = \{xG, P_m + xP_B\}$$

And to decrypt a cipher text, R1 multiplies the first point in the pair with R1's secret key and subtracts the result. It is represented by

$$P_m + xP_B - n_B(xG)$$
$$= P_m + x(n_B G) - n_B(xG) = P_m$$

Since the first sender has masked the message Pm, except S1 none can know the value of x. Though second sender knows P_B value, it cannot get the full message. For an attacker to get information, computation of x given G and xG is very difficult. Hence this method is very secure. The entire framework is composed of three level architecture with lower level containing the data bases and OLTP, middle level containing the architecture of data warehouse and security mechanisms in the higher level. Each level of the architecture is independent of the other with respect to the process of execution and this is the biggest advantage.

V. EXPERIMENTAL RESULTS AND DISCUSSION

Elliptic curve cryptography system calculates the multiplicative inverse of an integer for a given prime number using extended Euclid algorithm and the intermediate terms are less that prime numbers. The data is transformed so that it can be extracted correctly by Extraction Transformation Load Module inside the data warehouse.

Algorithm: transformation
Input I, N
I is the vector and N is the noise
for each attribute A_k
select the noise term e_k in N
$op_k \leftarrow$ {Multiply}
for each $I_i \in I$
do
for each a_k in $I_i = I$
(a_k is the observation of k^{th} attribute)
$a'_k \leftarrow \text{transform}(a_k, op_k, e_k)$
Output I'
End transform

The results of the above algorithm are used to perform association rule mining on the individual data. Series of experiments are conducted on medical data sets. The results are presented in the below table after applying our technique to the original databases for association rule mining process.

Table 1. Experimental Results in terms of accuracy

No. of Data Records	Using Original Data Set	Our method
500	89%	83%
1000	91%	87%
1500	91%	88%
2000	94%	90%
2500	93%	90%
3000	91%	89%
3500	89%	86%
4000	92%	88%

The above table highlights the accuracy of the association rule mining process conducted using the actual data and perturbed data. It is quite clear that the differences in accuracy are meager and our approach gives better privacy. It securely handles large volume of transmission between distributed data sets and data warehouse thereby minimizing the unauthorized access to the confidential data. Hence it is concluded that this method exhibits good privacy.

VI. CONCLUSION

Based on the experimental results and theoretical analysis, it can be concluded that the utilization of ECC for privacy preservation in association rule mining process can address the security challenges. The proposed method also reduces computing time. The use of ECC encryption in the PPARM applications protects the data or information from being accessed by unauthorized entities.

The performance of this method is purely dependent on selecting a proper elliptic curve for implementation. This is a distributed database combining scheme and uses data distortion technique for secure transmission of data from data sets to data warehouse. This method gives better security and privacy of data compared to other traditional approaches. In future, the same method is modified and implemented to test the efficiency by varying elliptic curve parameters.

VII. REFERENCES

[1]. Chen, CL Philip, and Chun-Yang Zhang. "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data." Information Sciences 275 (2014): 314-347.

[2]. Einav, Liran, and Jonathan Levin. "Economics in the age of big data." Science 346, no. 6210 (2014): 1243089.

[3]. Padhy, Neelamadhab, Dr Mishra, and Rasmita Panigrahi. "The survey of data mining applications and feature scope." arXiv preprint arXiv:1211.5723 (2012).

[4]. Saleem, Moazzam Ali, and Jehanzeb Hameed. "Integration of Data Mining and Object-Relational Database Systems." Ghulam Ishaq Khan Institute of Engineering Science and Technology, Swabi, NWFP, Pakistan (2003).

[5]. Rajan, J., and V. Saravanan. "A framework of an automated data mining system using autonomous intelligent agents." In Computer Science and Information Technology, 2008. ICCSIT'08. International Conference on, pp. 700-704. IEEE, 2008.

[6]. Wu, Xindong, Xingquan Zhu, Gong-Qing Wu, and Wei Ding. "Data mining with big data." IEEE transactions on knowledge and data engineering 26, no. 1 (2014): 97-107.

[7]. Appari, Ajit, and M. Eric Johnson. "Information security and privacy in healthcare: current state of research." International journal of Internet and enterprise management 6, no. 4 (2010): 279-314.

[8]. Zhu, Jian-Ming, Ning Zhang, and Zhan-Yu Li. "A new privacy preserving association rule mining algorithm based on hybrid partial hiding strategy." Cybernetics and Information Technologies 13, no. Special Issue (2013):41-50.

[9]. Dawid, Herbert. Adaptive learning by genetic algorithms: Analytical results and applications to economic models. Springer Science & Business Media, 2011.

[10]. Sinha, Tanmay, Vrns Srikanth, Mangal Sain, and Hoon Jae Lee. "Trends and research directions for privacy preserving approaches on the cloud." In Proceedings of the 6th ACM India Computing Convention, p. 21. ACM, 2013.

[11]. Shi, Yuliang, Zhongmin Zhou, Lizhen Cui, and Shijun Liu. "A sub chunk-confusion based privacy protection mechanism for association rules in cloud services." International Journal of Software Engineering and Knowledge Engineering 26, no. 04 (2016): 539-562.

[12]. Fan, Long. "Software and Website Development for Data Analysis and Management of DNA Barcoding." PhD diss., The Chinese University of Hong Kong (Hong Kong), 2014.

[13]. Rabiner, Lawrence R. "A tutorial on hidden Markov models and selected applications in speech recognition." Proceedings of the IEEE 77, no. 2 (1989): 257-286.

[14]. Chebrolu, Srilatha, Ajith Abraham, and Johnson P. Thomas. "Feature deduction and ensemble design of intrusion detection systems." Computers & security 24, no. 4 (2005): 295-307.

[15]. K. Iqbal, S. Asghar, and S. Fong, "A PPDM model using Bayesian Network for hiding sensitive XML Association Rules," in Digital Information Management (ICDIM), 2011 Sixth International Conference on, Melbourne, Australia, 2011, pp. 30-35.

[16]. J. Lai, Y. Li, R. H. Deng, J. Weng, C. Guan, and Q. Yan, "Towards semantically secure outsourcing of association rule mining on categorical data," Information Sciences, vol. 267, pp. 267-286, 5/20/2014.

[17]. Bernstein, Daniel J. "Introduction to post-quantum cryptography." In Post-quantum cryptography, pp. 1-14. Springer, Berlin, Heidelberg, 2009.

[18]. Caelli, William J., Edward P. Dawson, and Scott A. Rea. "PKI, elliptic curve cryptography, and digital signatures." Computers & Security 18, no. 1 (1999): 47-66.

[19]. Sari, Riri Fitri. "Selecting key generating elliptic curves for Privacy Preserving Association Rule Mining (PPARM)." In Wireless and Mobile (APWiMob), 2015 IEEE Asia Pacific Conference on, pp. 72-77. IEEE, 2015.