Computer Engineering and Applications Vol. 02, No. 03, December 2013

Two Phase Privacy Preserving Data Mining

Pooja Gupta[1], Ashish Kumar[2]

[1]*Department of Computer Science and Engineering, IFTM University Moradabad*
[2]*Department of Computer Science and Engineering, ITS Engineering College, Greater Noida*
[1]*say2poojagupta@gmail.com,* [2]*ashishcse29@gmail.com*

ABSTRAKSI

Makalah ini mengusulkan sebuah kerangka kerja untuk meningkatkan privasi data mining. Pendekatan yang digunakan memberikan keamanan di kedua ujungnya yaitu pada saat pengiriman data serta dalam proses data mining menggunakan dua tahap. Transmisi data yang aman ditangani menggunakan Elliptic Curve Cryptography (ECC) dan privasi yang dijaga menggunakan K-anonymity. Kerangka yang diusulkan menjamin lingkungan yang sangat aman. Kami mengamati bahwa kerangka melebihi pendekatan lain [8] dibahas dalam literatur di kedua ujungnya yaitu pada keamanan dan privasi data. Karena sebagian besar pendekatan telah dianggap baik transmisi aman atau privasi melestarikan data mining tetapi sangat sedikit yang dianggap baik. Kami telah menggunakan WEKA 3.6.9 untuk percobaan dan analisis pendekatan kami. Kami juga telah menganalisis kasus K-anonimity ketika jumlah catatan dalam kelompok kurang dari k (factor sembunyi) dengan menyisipkan catatan palsu. Hasil yang diperoleh menunjukkan pola bahwa penyisipan catatan palsu menyebabkan lebih akurat dibandingkan dengan penekanan penuh catatan. Sejak, penekanan penuh dapat menyembunyikan informasi penting dalam kasus di mana catatan kurang dari k, di sisi lain dalam proses catatan palsu penyisipan; catatan yang tersedia bahkan jika jumlah record dalam kelompok kurang dari k.

Kata kunci: K-anonymity, Rekaman Palsu, Eliptik, Kriptografi Kurva, Penggalian Data
Penjagaan Privasi

ABSTRACT

The paper proposes a framework to improve the privacy preserving data mining. The approach adopted provides security at both the ends i.e. at the data transmission time as well as in the data mining process using two phases. The secure data transmission is handled using elliptic curve cryptography (ECC) and the privacy is preserved using k-anonymity. The proposed framework ensures highly secure environment. We observed that the framework outperforms other approaches [8] discussed in the literature at both ends i.e. at security and privacy of data. Since most of the approaches have considered either secure transmission or privacy preserving data mining but very few have considered both. We have used WEKA 3.6.9 for experimentation and analysis of our approach. We have also analyzed the case of k-anonymity when the numbers of records in a group are less than k (hiding factor) by inserting fake records. The obtained results have shown the pattern that the insertion of fake records leads to more accuracy as compared to full suppression of records. Since, full suppression may hide important information in cases where records are less than k, on the other hand in the process of fake records insertion; records are available even if number of records in a group is less than k.

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

Keywords: K-anonymity, fake records, elliptic curve cryptography, privacy preserving data mining

3. INTRODUCTION

Nowadays most of our daily activities are routinely recorded and analyzed by variety of governmental and commercial organizations for the purpose of security and business related applications. Some of them are telephone calls, credit card purchases, internet surfing and sometimes even our medical records also. Such information might disclose an individual's privacy. Privacy has started gaining attention since 2000, because tremendous change has happened in the technologies since then. Also, the exponential growth of internet resulted in huge amount of data being stored in different types of databases on different sites. The privacy can be compromised while transferring data from various data sources to data warehouse, therefore, this end should also be secure. The aim of this paper is to make privacy preserving data mining secure. In our proposed approach we provided security at both the ends at the time of gathering the data from various sources and also at the time of mining the data. We have used ECC for secure transmission and k-anonymity for privacy preserving data mining.

4. RELATED WORK

Authors in [9] provided the concept of generalized table and minimal generalization. In [10] author discussed an algorithm *MinGen* combining the generalization and suppression to achieve k-anonymity.

ECC[5] is a public key cryptography approach based on the elliptic curves. Recently, ECC has gained attention since it uses smaller key size than its peer systems like RSA and DSA and maintains equivalent security level. The author in [7] has indicated that ECC is a promising area for exploration due to memory, bandwidth and computational constraints, and also, such areas demand for compaction of the devices and load balancing for environment. The [2] proposed a framework for privacy preserving data mining for distributed database**.** According to the framework [2] databases are encrypted using ECC before sending to the warehouse and are perturbed using multiplicative data perturbation (MDP) after decryption at the destination. Although the combination of ECC and MDP provides a strong security, however there are scenarios where perturbed data cannot be used for many data mining applications. Also, in [6] the objective of perturbation schemes is to mask the private data and still allowing summary statistics to be estimated. However, data mining techniques, such as clustering, classification, prediction and association rule mining, are necessarily need relationships among data records in order to provide correct results, not only summary statistic.

The privacy and security in data mining can be leaked mainly at two points (transmission and data mining). Therefore, in order to preserve privacy and security in data mining we should provide a strong framework that can ensure possible privacy leakage points. The model given in [10] provides k-anonymity protection which works with data mining, but does not consider security in data transmission. Most of the research works in this area attempt either considering privacy preservation in data mining [1, 4] or in secure transmission [3, 5]. However in [2] authors consider both privacy preserving data mining and secure transmission, but used the

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online

multiplicative perturbation which distorts each data element independently, therefore, Euclidean distance and inner product among data records are usually not preserved, so the perturbed data cannot be used for many data mining applications. In contrast the paper presents a framework which uses ECC for secure transmission and k-anonymity for PPDM.

5. PROPOSED EMBEDDED FRAMEWORK

The first phase of the developed integrated framework concerns with the identification of data sources and encryption using ECC before being transferred to data warehouse. Then the decryption of database is done to make it prepared for transformation. In the second phase k-anonymization is used for PPDM. This phase ensures the privacy of sensitive information.

3.1 DATESET INFORMATION

The sample database is a medical database for experiment. The dataset consists of 462 records and 25 attributes of patients. The generalization is applied on date of birth (DOB) and ZIP attributes, because these are quasi-identifiers and so can disclose privacy of patients.

3.2 METHODOLOGY

Phase-I (Secure Transmission Using ECC)

ECC uses a point G on an elliptic curve E over a finite field Z_P where p is a very large prime number [11]. D_A and D_B are databases which are distributed across the world. Assume that the sender S_A wants to send the database D_A to receiver R. Public and private key pair of sender S_A and receiver R is (P_A, K_A) and (P_R, K_R).P_{DA} is a point for database.

Step 1: *S_A and R agree on base point, G.*
Step 2: *Compute public and private keys.*
$P_A = K_A \cdot G$
$P_R = K_R \cdot G$
Step 3: *S_A selects a point P_{DA} on elliptic curve E.*
Step 4: *S_A encrypt the database D_A*
$E_{KA}(P_{DA}, K_A) = (K_A G, P_{DA} + K_A P_R)$
$= (y_1, y_2)$
Step 5: *R decrypts the database*
$D_{KR}(y_1, y_2) = y_2 - K_R y_1$
$= (P_{DA} + K_A P_R) - K_R (K_A G)$
$= P_{DA} + K_A (K_R G) - K_R (K_A G)$
$= P_{DA}$

Algorithm 1 *Phase-I Elliptic Curve Cryptography*

Input: D_A, k
Output: D_A' *(Anonymized Database)*
Step 1: *Find G_L*
Step 2: *Insert fake records in G_L*
(To satisfy k-anonymity)
Step 3: *Find G_{SI}*
Step 4: *Insert fake records in G_{SI}*
(With same values for QI and different values for SI)
Step 5: *Generalized of DOB and ZIP attributes.*

Algorithm 2 Phase-II k-Anonymity

FIGURE 1: Algorithms

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

Phase-II (Privacy Preservation Using K-Anonymity)
After the phase-I database D_A is securely retrieved by receiver R, now k-anonymity PPDM can be apply. Assume G_L is a group having number of records less than k and G_{SI} is a group having same sensitive information. QI and SI represent quasi-identifiers and sensitive information. The k-anonymity algorithm is:

The result of this phase ensures privacy of patient's record and is used for data mining. We have used Predictive *Apriori* technique for mining. The algorithms for both the phases are given in *Figure 1*.

4 RESULTS AND ANALYSIS

To test the performance of our proposed framework we performed a series of experiments on our sample database using **WEKA 3.6.9** tool. The experiments assumed different values of k = {4, 8, 12} and different cases. The *table 1*, *2* and *3* shows the sample of the results obtained in our experiments.

Table 1 Performance Table with Different k values

No. of Records	Accuracy of Results in %			
	Original Database	Anonymized Database (In case when there is no need to insertion of fake records)		
		K=4	K=8	K=12
200	87%	78%	75%	63%
462	88%	80%	76%	65%

Table 2 Performance without Fake Record Insertion

No. of Records	Accuracy Of Results in % (In case where no. of records in a group are less than k and we have suppressed those records)		
	K=4	K=8	K=12
200	60%	58%	50%
462	61%	59%	53%

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online

Table 3 Performance with Fake Record Insertion

No. of Records	Accuracy Of Results in % (In case where no. of records in a group are less than k and we have inserted fake records)		
	K=4	K=8	K=12
200	79%	77%	63%
462	82%	79%	64%

The results shown in *Figure 2* and *3* depicts that the insertion of fake records leads to more accuracy as compared to full suppression of records. Since, full suppression may hide important information in the cases where records are less than the values of k, on the other hand in the process of fake records insertion; records are available even if number of records in a group is less than k. The results obtained are promising and strengthen the hypothesis that proposed approach improves the privacy.

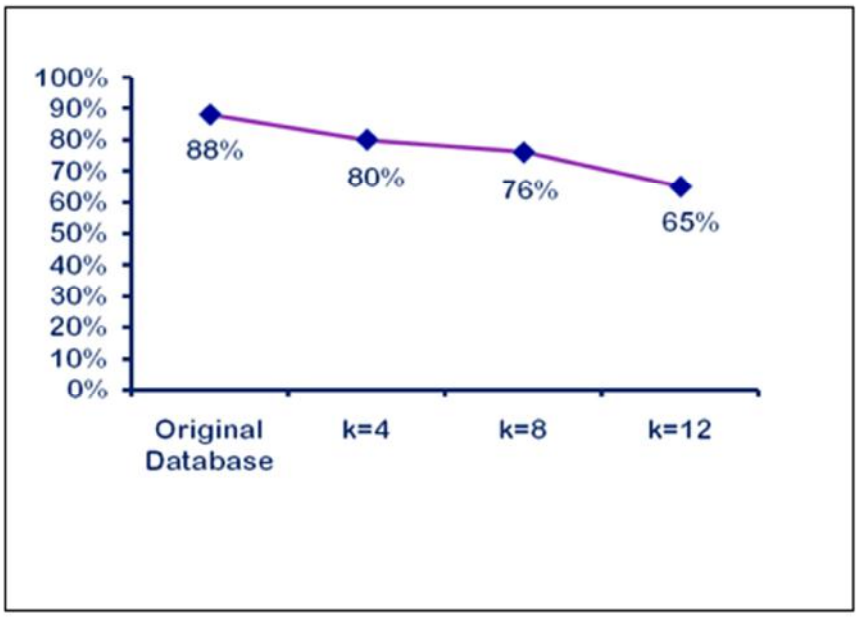

Figure 2 A Graph of k Vs Accuracy

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

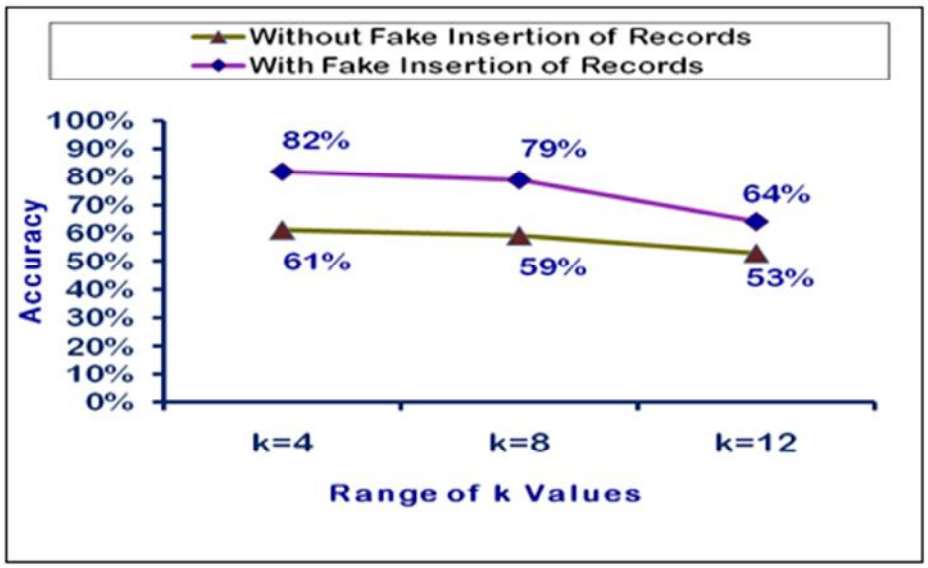

Figure 3 A Graph of k Vs Accuracy

5 CONCLUSION AND FUTURE DIRECTIONS

The wide proliferation to track and collect the databases distributed across different locations requires security. Databases can be distributed all over the world and can transmit to different locations. This leads to the need of security over communication as well as can preserve the privacy of individuals. We believe that the privacy of individuals is more important but the accuracy in results is also important but a small amount (negligible) of influence can be tolerated. We observed that k-anonymity algorithm does not reveal the sensitive information even after a group has the same sensitive information. Moreover the approach works well other data mining techniques.

It is also a good idea to extend our approach to handle privacy preservation of data streams. In cases of data streams, the privacy preservation is a bit challenging because the data is being released incrementally. In the proposed framework we have used Predictive Apriori data mining technique which belongs to Association data mining so, there is a need to design a PPDM mechanism for other data mining algorithms like classification, clustering and decision trees etc.

Future research may include the experimentation of our proposed framework on actual database released by any hospital and analyzing communication speed and overhead involved due to integration of ECC and PPDM. Performance can also be analyze for different cases such as a databases having records less than the k factor and also with all the records in a group of k having similar sensitive information.

REFERENCES

[1] Bayardo, R. J., and Agrawal, R., 'Data privacy through optimal k-anonymization' in Data Engineering, 2005. ICDE 2005: proc. of the 21st International Conference on pp. 217-228, 2005.

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online

[2] Kiran, P., Kumar, S. S., & Kavya, N. P., 'A Novel Framework using Elliptic Curve Cryptography for Extremely Secure Transmission in Distributed Privacy Preserving Data Mining', Advanced Computing: An International Journal, Vol.3, No.2, 2012.

[3] Malik, M. Y., 'Efficient implementation of elliptic curve cryptography using low-power digital signal processor', in Advanced Communication Technology (ICACT), 2010 The 12th International Conference on Vol. 2, pp. 1464-1468. IEEE, 2010.

[4] Mascetti, S., Bettini, C., Wang, X. S., & Jajodia, S., 'k-Anonymity in databases with timestamped data', in Temporal Representation and Reasoning, TIME, Thirteenth International Symposium on pp. 177-186, IEEE, 2006.

[5] Miller, V. S., 'Use of elliptic curves in cryptography', In Advances in Cryptology—CRYPTO'85 Proceedings, pp. 417-426 Springer Berlin Heidelberg, 1986.

[6] Pandya, B., Singh, U. K., Bunkar, K., & Dixit, K., "An Overview of Traditional Multiplicative Data Perturbation", International Journal, 2(3), 2012.

[7] Pateriya, R. K., & Vasudevan, S., 'Elliptic Curve Cryptography in Constrained Environments: A Review', In Communication Systems and Network Technologies (CSNT), International Conference on, pp. 120-124. IEEE, 2011.

[8] Samarati, P., 'Protecting respondents identities in microdata release', Knowledge and Data Engineering, IEEE Transactions on, 13(6), 1010-1027, 2001.

[9] Sweeney, L. & Samarati, P., 'Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression' Technical report, SRI International, 1998.

[10] Sweeney, L., "Achieving k-anonymity privacy protection using generalization and suppression", International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05), 571-588, 2002.

[11] Hankerson, D., Vanstone, S., & Menezes, A. J., "Guide to elliptic curve cryptography", Springer, 2006.

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online