World Engineering & Applied Sciences Journal 7 (3): 190-193, 2016
ISSN 2079-2204

DOI: 10.5829/idosi.weasj.2016.7.3.22525

A Secure Scheme for Privacy Preserving Data Mining Using Matrix Encoding

[1]J. Shana and [2]T. Venkatachalam

[1]Department of Computer Application, Coimbatore Institute of Technology, Coimbatore, India
[2]Department of Physics, Coimbatore Institute of Technology, Coimbatore, India

Abstract: Organizations are generating data in large amount which is being subjected to analysis to extract useful information out of them. Also there is multiple storages on disks and transfer of data before applying data mining algorithms on them. Preserving the privacy of the dataset as well the mined results has become very important for the data owners. In this paper we suggest a novel scheme to protect the data at rest before it moves for analysis. It uses a special form of encoding using the concept of matrix theory. The encoded data matrix is assessable to only the authorized data miner who has the key for the encoder matrix. The goal is to protect the dataset without affecting the mining results. Experiments were conducted on synthetic and real datasets and it proves that this simple method is effective in preserving the privacy of data against the usual methods of data encryption.

Key words: Privacy · Data mining · Matrix theory · Encryption

INTRODUCTION

Huge amount of data is generated in various forms in organizations. Many of them are into in-house data analytics where data is subjected to various analysis techniques to discover hidden knowledge. Data get stored on disks, moved to back-ups or uploaded to servers for analysis. It then becomes necessary to protect the dataset to preserve its privacy. Data mining discovers hidden patterns from large datasets by applying techniques and methods from areas of machine learning, computer science, mathematics and statistics [1]. Explosive progress in networking, storage and processor technology has led to the creation of ultra large databases that record unprecedented amount of transactional information. The main problem is that with the availability of non-sensitive information or unclassified data, one is able to infer sensitive information that is not supposed to be disclosed. Despite its benefits in various areas such as marketing, business, medical analysis, bioinformatics and others, data mining can also pose a threat to privacy in database security if not done or used properly. Privacy preserving data mining, is a novel research direction in data mining and statistical databases, where data mining algorithms are analyzed for the side-effects they incur in data privacy. There are many approaches followed by researches for privacy preserved data mining. A detailed reference can be obtained from [2]. This paper focuses on protecting the database before it goes for mining in a centralized environment. In paper we suggest a novel scheme for encoding the database using matrix termed as matrix based encoding. It can be very well extended to a distributed environment also. Rest of the paper is divided into following sections. Section

Literature Survey: Many researchers proposed many methods for privacy preserving mining for both centralized and distributed databases. The state of the art in the area of privacy preserving data mining techniques is discussed by the authors in [3]. This paper also describes the different dimensions of preserving data mining techniques such as data distribution, data modification technique, data mining algorithms, data or rule hiding and approaches for privacy preserving data mining techniques. The survey of basic paradigms and notations of secure multiparty computations and their relevance to the field of privacy preserving data mining are presented by the authors in [4]. They also discussed the issue of efficiency and demonstrate the difficulties involved in constructing highly efficient protocols. In [5], the authors proposed a framework for evaluating privacy preserving data mining algorithms and based on their

Corresponding Author: J. Shana, Department of Computer Application, Coimbatore Institute of Technology, Coimbatore, India.

frame work one can assess the different features of privacy preserving algorithms according to different evaluation criteria. In [6], Secure mining of association rules over horizontally partitioned database using cryptographic technique to minimize the information shared by adding the overhead to the mining process is presented. In [7], authors proposed an enhanced scheme two- phase privacy preserving in distributed data mining. They presented two protocols to increase the security against collusion in the communication environment with or without trusted party. A new algorithm which is the modification of the existing algorithm and based on a semi-honest model with negligible collision probability is proposed in [8]. They also used cryptography techniques to preserve the privacy. Privacy preserving in data mining by using cryptographic role based access control is presented in [9]. They proposed a new solution by integrating the advantages of the first approach which protects the privacy of the data by using an extended role based access control approach and the second approach which uses cryptographic techniques with the view of minimizing loss of information and privacy. In [10], authors addresses the problem of association rule mining in vertically partitioned database by using cryptography based approach and also presents the analysis of security and communication.

Proposed Scheme

Architecture: In this paper we propose a novel scheme to protect the dataset. Figure 1 shows the architecture of the proposed system. There are two phases in this model. First phase is termed the Encoding Phase where the data owner encodes the dataset. The second phase is the Decoding Phase, where only the authorized data miner can perform the decoding of the dataset and apply the mining algorithm [11].

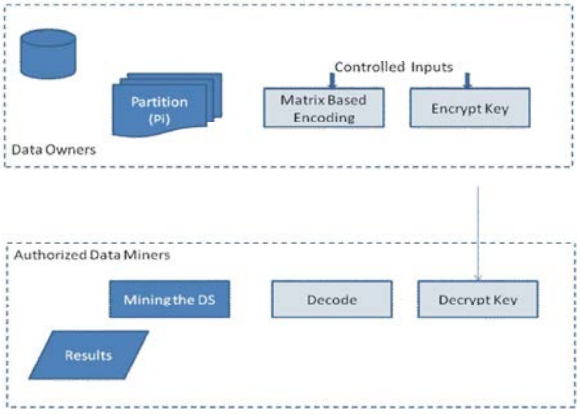

Fig. 1: Architecture of Proposed System

Table 1: Algorithm for Matrix encoding and decoding

Input: {D_i i = 1 to num_partitions}
Output= {X_i } // encoded matrices
Encoding
1: Choose a Quadratic Form(QF) of n variables to form the encoded matrix
2: Derive the induced diagonal matrix from QF, E.
2: Encode as X= E.D //matrix multiplication
3: Encrypt the QF ///key
Decoding
4: Decrypt the QF
5: Identify the matrix and its inverse
6: Perform D= $X.E^{-1}$
7. Apply mining algorithm on D

Encoding Phase: Here the dataset is first partitioned into sub-datasets. Any kind of partition horizontal or vertical can be applied to reduce the dataset into manageable units according to the memory and computation resources available. Then each partition is converted into a matrix form using any numerical encoding method. Table 1 shows the algorithmic steps for matrix encoding used in this scheme.

Here an arbitrary non-singular matrix is chosen for encoding the given dataset. For example, consider a matrix.

$$D = \begin{pmatrix} 5 & 0 & -4 \\ 0 & -11 & 5 \\ -6 & 5 & -7 \\ 8 & -7 & 5 \\ -11 & 8 & 0 \\ 3 & -2 & 0 \end{pmatrix}$$

An arbitrary non-singular matrix A is chosen for decoding the data matrix. The inverse of which is used for decoding.

$$A = \begin{pmatrix} -1 & 5 & -1 \\ -2 & 11 & 7 \\ 1 & -5 & 2 \end{pmatrix} \text{ whose inverse is given by } A^{-1} = \begin{pmatrix} 57 & -5 & 46 \\ 11 & -1 & 9 \\ -1 & 0 & -1 \end{pmatrix}$$

Decoding Phase: In the decoding phase, the inverse of matrix is used to derive the original dataset as described below.

$$X = DA = \begin{pmatrix} 5 & 0 & -4 \\ 0 & -11 & 5 \\ -6 & 5 & -7 \\ 8 & -7 & 5 \\ -11 & 8 & 0 \\ 3 & -2 & 0 \end{pmatrix} \begin{pmatrix} -1 & 5 & -1 \\ -2 & 11 & 7 \\ 1 & -5 & 2 \end{pmatrix}$$

$$= \begin{pmatrix} -9 & 45 & -13 \\ 27 & -146 & -67 \\ -11 & 60 & 27 \\ 11 & -62 & -47 \\ -5 & 33 & 67 \\ 1 & -7 & -17 \end{pmatrix}$$

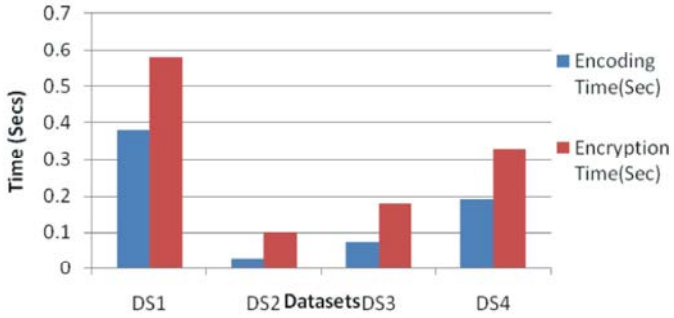

Fig. 2: Dataset vs encoding Time

Table 2: Dataset Characteristics

Dataset	#Distinct Items	#N	T_{max}
TD100K	1000	100000	31
TD1000	1000	1000	19
BMSWebView1	267	59602	267
BMSWebView2	3340	77512	161

Table 3: Storage requirements for the partitions

Dataset	m	N	#P_Size	Memory/Partition (KB)	Total Memory(MB)
DS 1	500	267	120	521.48	61.11
DS 4	500	161	155	314.45	47.59
DS 3	500	31	200	60.54	11.82
DS 2	500	19	1	37.11	0.072

Experiments and Results: The characteristics of the dataset used are shown in Table 2. Here two synthetic and two real datasets from FIMI repository are used for studying the encoding time of the data matrix. The experiments were focused on studying the encoding time for the dataset. The storage requirements were also studied using the horizontal and vertical partitions. All the experiments were run on a 4GB machine with 400GB hard disk running Windows 7 Home Premium Edition. The proposed method is tested against the known content based encryption adopted in all database servers. Figure 1 shows the time taken for encoding each partition using the matrix based approach and the usual content encryption. There is significant difference decrease in time for the proposed method because matrix operation in memory is faster compared to encryption and decryption. Also the storage requirements in terms of partition size are manageable as shown in Table 3.

CONCLUSION

The experimental results prove that the time taken for encoding is significantly less for larger matrices. Also the two way protection scheme enables to protect the dataset from intruders. There is no one standard metric to quantify privacy. Here we assume the adversary- knowledge model. What knowledge will an intruder possess in order to gain access of the original dataset? Anytime the dataset is in encoded form. Let us assume that the intruder is able to gain access to the encoder matrix. This is possible only if the intruder is able to know the matrix itself. Here we encrypt the matrix form and access of the decryption key is given to only authorized persons. This method is a simple effective solution for protecting the dataset for data mining. As further extension we are to study how the model behaves when big data in distributed environment is given as input.

REFERENCES

1. Han, Jiawei and Micheline Kamber, 2006. Data Mining Concepts and Techniques, 2006, 2nd edition, Morgan Kaufmann Publishers, ISBN 155860-901-6.
2. Agrawal, R. and R. Srikant, 2000. Privacy-preserving data mining, Proc of the ACM SIGMOD, Intl Conf. on Management of Data, Dallas, Texas, pp: 439-450.
3. Agrawal, D. and C. Aggarwal, 2001. On the Design and Quantification of Privacy Preserving Data Mining Algorithms. Proceedings of PODS, 21(2): 247-255.

4. Verykios, V.S., E. Bertino, I. Nai Fovino, L. Parasiliti, Y. Saygin and Y. Theodoridis, 2004. State- of-the-art in privacy preserving data mining. SIGMOD Record, 33(1): 50-57.
5. Lindell, Y. and B. Pinkas, 2009. Secure Multiparty Computation for Privacy-Preserving Data Mining, The Journal of Privacy and Confidentiality, 1: 59-98.
6. Bertino Elisa , Igor Nai Fovino Loredana and Parasiliti Provenza, 2005. A Framework for Evaluating Privacy Preserving Data Mining Algorithms, Data Mining and Knowledge Discovery, 11: 121-154.
7. Kantarcioglu, M. and C. Clifto, 2002. Privacy-preserving distributed mining of association rules on horizontally partitioned data, In IEEETransactions on Knowledge and Data Engineering Journal, 16(9): 1026-1037.
8. Chin-Chen Chang, Jieh-Shan Yeh and Yu-Chiang Li, 2006. Privacy- Preserving Mining of Association Rules on DistributedDatabases, IJCSNS International Journal of Computer Science and Network Security, 6(11).
9. Mahmoud Hussein, Ashraf El-Sisi and Nabil Ismail, 2008. Fast Cryptographic Privacy Preserving Association Rules Mining on Distributed Homogenous Data Base, I. Lovrek, R.J. Howlett and L.C. Jain (Eds.), LNAI 5178, Springer-Verlag Berlin Heidelberg, pp: 607-616.
10. Vasudevan Lalanthika, S.E. Deepa Sukanya and N. Aarthi , 2008. Privacy Preserving Data Mining Using Cryptographic Role Based Access Control Approach, Proceedings of the International Multi Conference of Engineers and Computer Scientists, 1: 61-70.
11. Vaidya, J. and C. Clifton, 2002. Privacy preserving association rule mining in vertically partitioned data, 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, pp: 639-644.