International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Protection of Private Data in Association Rule Mining

Ashwini B. Yadav[1], Vikas B. Maral[2]

[1]Savitribai Phule Pune University, K. J. College of Engineering & Management Research, India
[2]Assistant Professor at K. J. College of Engineering & Management Research, Savitribai Phule Pune University, India

Abstract: ***Today, in the computerized world there is huge requirement of storing data very securely. So this project is mainly based on providing security to the organizations data on the outsourced databases. For example, in cloud computing an organization which lacks in expertise and computing services can fulfil its mining needs from service provider. So to make company's private data protected that company which is data owner, ships data in encrypted form to third party service provider. And then sends mining queries to service provider to returns pattern with true support. To achieve this, encrypt/decrypt(E/D) module is used. Firstly, data set at client side is encoded and again encryption is applied on encoded data. To transform encoded data in encrypted form, an algorithm is used named as AES algorithm under the encrypt/decrypt scheme. Under AES algorithm, Diffie-Hellman algorithm is used key generation. Once key is generated, AES is used to encrypt and decrypt the data. E/D scheme is used for avoiding service provider from sharing data to other parties other than the data owner. Now the service provider is unaware about what data is stored on server.***

Keywords: association rules, service provider, data owner, encryption, decryption.

1. Introduction

Cloud computing is the long dreamed vision of computing as a utility, where cloud customers can remotely store their data into the cloud so as to enjoy the on-demand high quality applications and services from a shared pool of configurable computing resources. Its great flexibility and economic savings are motivating both individuals and enterprises to outsource their local complex data management system into the cloud, especially when the data produced by them that need to be stored and utilized is rapidly increasing.

In outsourcing data to server or before disclosing the database, sensitive patterns must be hidden to enhance the security of database. For example Let a clothes store that purchase jeans from two companies, A and B, and both can access customers' database of the store. Now A applies data mining techniques and mines association rules related to B's products. A had found that most of the customer who buy jeans of the B also buy belt. Now A offers some discount on belt if customer purchases A's jeans. As result the business of B goes down. So releasing the database with sensitive information cause the problem. This scenario gives the direction to research on sensitive rules (or knowledge) hiding in database.

Association rules are if/then statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository. An example of an association rule would be "If a customer buys a dozen eggs, he is 80% likely to also purchase milk." Association rules are created by analyzing data for frequent if/then patterns and using the criteria support and confidence to identify the most important relationships.

2. Existing System

In the field of data mining, privacy protection is very important issue. Previous researches on association rule mining focuses on protection of sensitive data from third party players only while protecting that data from service provider is also an important issue [1,5,6].

For the purpose of encryption of transactional data, a substitution cipher method is used in [3].A Rob Frugal algorithm is used for encryption before outsourcing the private data [4].RSA encryption algorithm is used by Tinghuai Ma, Sainan Wang [2].It has drawback of the ciphertext generated by this method is 8 times longer than plaintext. In proposed system, we are using AES encryption algorithm which is very powerful encryption technique.

3. Proposed System

In the proposed system, we basically provide security to data owner's private data. In this scheme, data owner firstly registers with service provider, then he can store local database to server side. But before outsourcing local database of client to server, it will be encoded in the binary format. Then AES encryption technique will be applied to this encoded data. The keys required for this technique will be generated by using Diffie Hellman algorithm. Here for encryption RSA encryption technique can be used, but its results are eight times longer than original data. So, we are going to use AES technique which is a powerful technique. Then after encryption, the encrypted data is outsourced to the server side. At server side, using decryption key, the encrypted data gets decrypted and results in encoded data. Here we are not performing decoding. A modified Apriori algorithm is applied to encoded data which finds association rules between data items which are in the encoded format. This result is given to the client side. At client side, the result in encoded format gets decoded and plaintext is generated.

4. Mathematical Model

Let s (be a main set of) ≡ {SDB, LDB, C, A, S, MR, AO}
where,

SDB is the copy of the server database. This database is responsible for storing user information related to cloud interactions.

LDB is a set of local database that a user owns. It consists of data tables having data items related to the products and their sales transactions.

C is a set of all clients using the server database and mining services from the server. And (c_1 , c_2 , c_3, ...c_n) Є C.

A is a set of algorithms applied on the input data to get mining results. S is the server component of the system. The server is responsible for registering, authenticating and providing associations to the end user.

MR is a set of mining rules that are applied on the input dataset provided by the client from his LDB. And (mr_1 , mr_2 , mr_3, ...mr_n) Є MR

AO is a set of associations that are extracted from the input and a form the output of the system.

Functionalities:
SDB' = RegisterUser(uid, password, fullname, address, country, contact, email);
password = SHA1(input_password);
U = AuthenticateUser(uid, password, SDB');
LDB1 = ManageProducts(pid, product name, cost);
LDB2 = ManageBilling(transactions, items);
LDB = LDB1 + LDB2
ED(Encoded data) = EncodeTransactions(LDB2, EncodingAlgorithm(EA));
UPLOAD(ED);
AO = Apply Mining(ED);
Results = Decode (Download (AO));

Table 1: Mathematical Model

Input	*Process*	*Output*
User Details	Registration	SDB'
UID, Password	Authentication	Valid User
LDB1,LDB2,Encoding Algo	Encoding	Encoded Data
Encoded Data, encrypt Key	Encryption	Encypted Data
[At server] Encrypted Data,Decrypt Key	Decryption	Decrypted Data
Decrypted Data	Mining, Decoding	Mining Results(decoded data)

5. Architecture

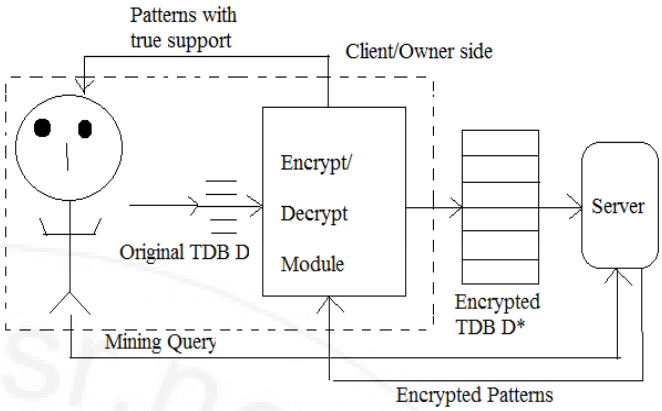

Figure 1: Architecture

System architecture shown in above figure represents the overall flow of the system. We can easily understand how selected database is encoded and encrypted, how the client gets final results by maintaining privacy of data throughout the system.

- **Client**
  Client is the entity in system which may be a data owner who stores large database to the cloud server to minimize complexity of handling it locally or he may be the user who requests for the association rules between itemsets to the sever.
- **Encrypt/Decrypt module**
  Encrypt/Decrypt module is the part of system where encryption and decryption of database is done to preserve the privacy of sensitive data. Firstly data owner who wants to store its database to the server will send data to encrypt/decrypt module. Then this module will encrypt database and stores it to the server.
  Now when a user fires a mining query to database stored at server, Apriori algorithm will generate association between different item sets and sends it to encrypt/decrypt module, where the results gets decrypted and returned to the user.
- **Server**
  Server is the service provider or cloud server which stores large databases of different data owners. Server provides different resources or services to the user. It is the entity where every user requests for different services.

6. Implementation Methods

In the proposed system, we first encode users sensitive data, then encryption and decryption processes are performed on that encoded data by using AES algorithm. So this algorithm requires key generation to implement it and the key is generated by using Diffie-Hellman algorithm. Diffie-Hellman is a key exchange algorithm. So this generated key is stored in server database for the further decryption process. Then Apriori algorithm is applied to data stored at server side that attempts to find subsets which are common to at least a minimum number C (the cutoff, or confidence threshold) of the itemsets.

6.1 AES

- The overall structure of AES encryption/decryption is shown in Figure 2.
- The number of rounds shown in Figure is 10, for the case when the encryption key is 128 bit long. (the number of rounds is 12 when the key is 192 bits, and 14 when the key is 256.)
- Before any round-based processing for encryption can begin, the input state array is XORed with the first four words of the key schedule. The same thing happens during decryption — except that now we XOR the ciphertext state array with the last four words of the key schedule.
- For encryption, each round consists of the following four steps:
- 1) Substitute bytes, 2) Shift rows, 3) Mix columns, and 4) Add round key. The last step consists of XORing the output of the previous three steps with four words from the key schedule.
- For decryption, each round consists of the following four steps: 1) Inverse shift rows, 2) inverse substitute bytes, 3) Add round key, and 4) Inverse mix columns. The third step consists of XORing the output of the previous two steps with four words from the key schedule
- The last round for encryption does not involve the “Mix columns” step. The last round for decryption does not involve the “Inverse mix columns” step.

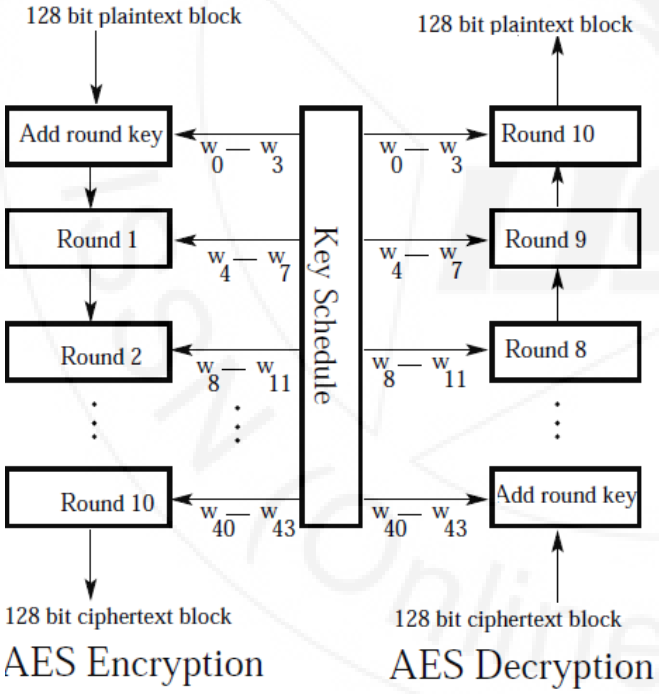

Figure 2: AES implementation

6.2 Apriori Algorithm

In Computer science and data mining, Apriori is classic algorithm for learning association rules. Apriori is designed to operate on databases containing transactions (for example, collections of items bought by customers, or details of a website frequentation). The algorithm attempts to find subsets which are common to at least a minimum number C (the cutoff, or confidence threshold) of the itemsets. Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time (a step known as *candidate generation*, and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found. Apriori uses breadth-first search and a hash tree structure to count candidate item sets efficiently.

6.3 Deffie-Hellman Algorithm

- Defffie-Hellman key agreement protocol was developed by Diffie and Hellman in 1976.This protocol allows two users to exchange a secrete key over an insecure medium without any prior secretes.The protocol has two system parameters P and g.They are both public and may be used by all users in a system.
- The parameter p is prime number and g is an integer less than p,with the following property that for every number n between 1 and p-1 inclusive there is a power k of g such that n=g^k mod p. The protocol depends on the descrete logarithms problem for its security.
- The Diffie-Hellman key exchange does not authenticate the participants.It is vulnerable to man-in-middle attack.
- Algorithm:
- Select two numbers: one prime number q and other integer number that is primitive root of q.
- Suppose the users A and B wants to exchange a key.

1. User A selects a random integer XA <q and computes YA=α^XA mod q.
2. User B selects a random integer XB<q and computes YB=α^XB mod q.
3. Both side keeps the X value private and makes Y value public to other side.
4. User A computes the key as K= (YB)^XA mod q
    1. User B computes the key as K= (YA)^XB mod q

- Both sides gets same results :

K= (YB)^XA mod q
=(α^XB mod q)^XA mod q
= (α^XB)^ XA mod q
=α^XB XA mod q
= (α^XA mod q)^XB mod q
= (YA)^XA mod q.

7. Conclusion

In this paper, we studied protection of outsourced data in association rule mining. The basic premise was the sensitive data is transformed in another format by using encrypt/decrypt scheme so that the sensitive data will be unreadable by anyone even by the semitrusted server.

References

[1] R. Buyya, C. S. Yeo, and S. Venugopal, “Market-oriented cloud computing: Vision, hype, and reality for delivering it services as computing utilities,” in *Proc. IEEE Conf. High Performance Comput. Commun.*, Sep. 2008, pp. 5–13.

[2] W. K. Wong, D. W. Cheung, E. Hung, B. Kao, and N. Mamoulis, "Security in outsourcing of association rule mining," in *Proc. Int. Conf.Very Large Data Bases*, 2007, pp. 111–122.
[3] L. Qiu, Y. Li, and X. Wu, "Protecting business intelligence and customer privacy while outsourcing data mining tasks," *Knowledge Inform. Syst.*, vol. 17, no. 1, pp. 99–120, 2008.
[4] C. Clifton, M. Kantarcioglu, and J. Vaidya, "Defining privacy for data mining," in *Proc. Nat. Sci. Found. Workshop Next Generation Data Mining*, 2002, pp. 126–133.
[5] I. Molloy, N. Li, and T. Li, "On the (in)security and (im)practicality of outsourcing precise association rule mining," in *Proc. IEEE Int. Conf. Data Mining*, Dec. 2009, pp. 872–877.
[6] F. Giannotti, L. V. Lakshmanan, A. Monreale, D. Pedreschi, and H. Wang, "Privacy-preserving data mining from outsourced databases," in *Proc. SPCC2010 Conjunction with CPDP*, 2010, pp. 411–426.
[7] R. Agrawal and R. Srikant, "Privacy-preserving data mining," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2000, pp. 439–450.
[8] S. J. Rizvi and J. R. Haritsa, "Maintaining data privacy in association rule mining," in *Proc. Int. Conf. Very Large Data Bases*, 2002, pp. 682– 693.
[9] M. Kantarcioglu and C. Clifton, "Privacy-preserving distributed mining of association rules on horizontally partitioned data," *IEEE Trans. Knowledge Data Eng.*, vol. 16, no. 9, pp. 1026–1037, Sep. 2004.
[10] B. Gilburd, A. Schuster, and R. Wolff, "k-ttp: A new privacy model for large scale distributed environments," in *Proc. Int. Conf. Very Large Data Bases*, 2005, pp. 563–568.

Author Profile

Ashwini B. Yadav Student of K. J. College of Engineering and Management Research, Pune, Maharashtra, India