International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Privacy preserving technique ECC for secure transmission in field of Data Mining

M.Mohanrao[1], Dr.S.Karthik[2]
[1]Research scholar Bharathiar University, Coimbatore, India [2]Professor & Dean / CSE, SNS College of Technology, Coimbatore

Abstract: ***This paper aims at bring forward Technique on distributed Data Mining as Data mining is used for retrieving intelligent information from enormous databases. Presently these databases are spread across the world. Tried to make a view that preserves the privacy in distributed mining. Approached with distributed data clustering algorithm. Mining algorithms methods and trends in order to discover knowledge from distributed data in accurate way. PPDM research usually uses: data hiding, in which sensitive raw data like identifiers, name, addresses, etc. were altered, blocked, or trimmed out from the original database, at the same time the users of the data not to be able to compromise with other person's confidentiality; and anther rule hiding, in which sensitive knowledge extracted from the data mining process be excluded for use, because confidential information may be Derived from the released knowledge. This problem is also commonly called the "database inference problem;" and lastly Secure Multiparty Computation (SMC), where distributed data are encrypted before released or shared for computations, Data encryption techniques that used in security.***

Keywords: ECC transmission, privacy, distribution, blocking, data mining, clusters

I. INTRODUCTION

Data mining is widely used by researchers for science and business purposes. Privacy Preserving Data Mining is a method which ensures privacy of individual information during mining. Most important task involves retrieving information from multiple data bases which is distributed. Distributed Data Mining (DDM) is concerned with the application of the classical Data Mining procedure in a distributed computing environment trying to make the best of the available resources. Data Mining takes place both locally at each distributed site and at a global level where the local knowledge is fused in order to discover global knowledge. The continuous developments in information and communication technology have recently led to the appearance of distributed computing environments, which comprise several, and different sources of large volumes of data and several computing units.

PPDM is a fast growing research area. Given the number of different algorithms have been developed over the last years, there is an emerging need of synthesizing literature to understand the nature of problem, identify potential research issues, standardize new research area, and evaluate the relative performance of different approaches. PPDM is a given the number of different algorithms have been developed over the last years, there is an emerging need of synthesizing literature to understand the nature of problem, identify potential research issues, standardize new research area, and evaluate the relative performance of different approaches.

The purpose of this paper is to propose a novel frame work which ensures both secure transmission and confidentiality of information since most of the earlier research has been done on either secure transmission or PPDM method. This frame work can be divided into data hiding, and encryption in which sensitive raw data like identifiers, name, addresses, etc. were altered, blocked, or trimmed out from the original database, in order for the users of the data not to be able to compromise another person's privacy; and anther rule hiding, in which sensitive knowledge extracted from the data mining process be excluded for use, because confidential information may be Derived from the released knowledge. This problem is also commonly called the "database inference problem;" and lastly Secure Multiparty Computation (SMC), we provide a cryptographic method for secure transmission using elliptic curve cryptography. Where distributed data are encrypted before released or shared for computations; thus, no party knows anything except its own inputs and the results.

II. RELATED WORK

The application of the classical knowledge discovery process in distributed environments requires the collection of distributed data in a data warehouse for central processing. However, this is usually either ineffective or infeasible for the following reasons:

A. Storage Cost

It is obvious that the requirements of a central storage system are enormous. A classical example Concerns data from the astronomy science, and especially images from earth and space telescopes. The size of Such databases are reaching the scales of exabytes (1018 bytes) and is increasing at a high pace. The central storage of the data of all telescopes of the planet would require a huge data warehouse of enormous cost.

B. Communication Cost

The transfer of huge data volumes over network might take extremely much time and also require an unbearable financial cost. Even a small volume of data might create problems in wireless network environments with limited bandwidth. Note also that communication may be a continuous overhead, as distributed databases are not always constant and unchangeable. On the contrary, it is common to have databases that are frequently updated with new data or data streams that constantly record information.

C. Private and Sensitive Data

There are many popular data mining applications that deal with sensitive data, such as people's medical and financial records. The central collection of such data is not desirable as it puts their privacy into risk. In certain cases (e.g. banking, telecommunication) the data might belong to different, perhaps competing, organizations that want to exchange knowledge without the exchange of raw private data

III. PROPOSED FRAMEWORK

Privacy preserving has originated as an important concern with reference to the success of the data mining. Privacy preserving data mining deals with protecting the privacy of individual data or sensitive knowledge without sacrificing the utility of the data. People have become well aware of the privacy intrusions on their personal data and are very reluctant to share their sensitive information. This may lead to the inadvertent results of the data mining. Within the constraints of privacy, several methods have been proposed but still this branch of research is in its infancy.

A. Process followed in this approach

This programming method very basically the data base is distributed in to several sub unites , as the data undergoes division, the basic theme behind making such approach; is to keep the data confidential, to keep the data secure and hide from manipulations. One of such method is Heuristic approach, i.e to keep the data hidden. Upon assigning some more rules the same data undergoes a process of Hierarchal clustering that again included with some more processing methods like Agglomerative, Hierarchal. At the final stage obtained data processed with ECC method for securing transmission.

B. Hiding Methodology

The PPDM algorithms are mainly used on the tasks of classification, association rule and clustering. Association analysis involves the discovery of associated rules, showing attribute value and conditions that occur frequently in a given set of data. Classification is the process of finding a set of models that describe and is sway data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. Clustering Analysis concerns the problem of decomposing or partitioning a data set (usually multivariate) into groups so that the points in one group are similar to each other and are as different as possible from the points in other groups. There are three types of Association rule hiding algorithms namely border-based approaches, exact approaches and heuristic approaches. Heuristic approaches are very efficient and fast algorithms that modify the selected transactions from the database for hiding the sensitive knowledge. There are two types of heuristic approaches exist namely data blocking method and data distortion method. This chapter describes about the types of heuristic approaches and compares the performance with the proposed weight based sorting distortion algorithm in terms of hiding failure and data quality.

Given a rule r and calculate minconf(r), maxconf(r) as

minconf(r) = minsup(r)*100 / maxsup (lr) ----------(1)

maxconf(r) = maxsup(r)*100 / minsup (lr) ---------- (2)

Where lr denotes the rule antecedent.

Considering the support interval and the minimum support threshold have the following cases for an itemset A:

(i) A is hidden when maxsup(A) is smaller than MST

(ii)A is visible with an uncertainty level when minsup (A) ≤ MST ≤ maxsup(A)
(iii) A is visible if minsup(A) is greater than or equal to MST

They showed the architecture of association rule hiding. A rule hiding process takes place according to two different strategies: decreasing its support or its confidence. In this method, the adopted alternative strategies aim at introducing uncertainty in the frequency or the importance of the rules to hide. The two strategies reduce the minimum support and the minimum confidence of the item sets generating these rules below the Minimum Support Threshold (MST) and Minimum Confidence Threshold (MCT) correspondingly by a certain Safety Margin Threshold (SMT) fixed by the user. In order to reduce the support of the large itemset generating a sensitive rule,

Algorithm 1 replaces 1"s by "? "For the items in transactions supporting the itemset until its minimum support goes below the minimum support threshold MST by the fixed safety margin SM.

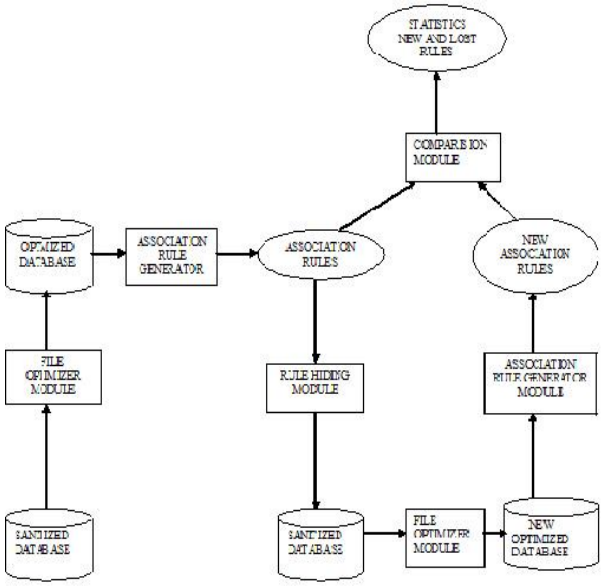

Figure: Architecture of Association Rule Hiding

The first rule decreases the minimum support of the generating item set of a sensitive rule by replacing items of the rule consequent with unknown values. Whereas the second rule increases the maximum support value of the antecedent of the rule to hide via placing question marks in the place of the zero values of items in the antecedent.
The PPDM algorithms are mainly used on the tasks of classification, association rule and clustering. Association analysis involves the discovery of associated rules, showing attribute value and conditions that occur frequently in a given set of data. Classification is the process of finding a set of models (or functions) that describe and istinguish data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. Clustering Analysis concerns the problem of decomposing or partitioning a data set (usually multivariate) into groups so that the points in one group are similar to each other and are as different as possible from the points in other groups.

C. Hierarchical clustering algorithm

Divisive Hierarchical clustering - It is just the reverse of Agglomerative Hierarchical approach.Hierarchical clustering algorithm is of two types:

i) Agglomerative Hierarchical clustering algorithm or AGNES (agglomerative nesting) and
ii) Divisive Hierarchical clustering algorithm or DIANA (divisive analysis).
Both this algorithm is exactly reverse of each other. So we will be covering Agglomerative Hierarchical clustering algorithm in detail.

Agglomerative Hierarchical clustering -This algorithm works by grouping the data one by one on the basis of the nearest distance measure of all the pair wise distance between the data point. Again distance between the data point is recalculated but which distance to consider when the groups has been formed? For this there are many available methods. Some of them are:
1) single-nearest distance or single linkage.
2) complete-farthest distance or complete linkage.
3) average-average distance or average linkage.
4) centroid distance.
5) ward's method - sum of squared Euclidean distance is minimized.
This way we go on grouping the data until one cluster is formed. Now on the basis of dendogram graph we can calculate how many numbers of clusters should be actually present.

D. Algorithmic steps for Agglomerative Hierarchical clustering

Let X = {x1, x2, x3, ..., xn} be the set of data points.

1) Begin with the disjoint clustering having level L(0) = 0 and sequence number m = 0.

2) Find the least distance pair of clusters in the current clustering, say pair (r), (s), according to d[(r),(s)] = min d[(i),(j)] where the minimum is over all pairs of clusters in the current clustering.

3) Increment the sequence number: m = m +1.Merge clusters (r) and (s) into a single cluster to form the next clustering m. Set the level of this clustering to L(m) = d[(r),(s)].
4) Update the distance matrix, D, by deleting the rows and columns corresponding to clusters (r) and (s) and adding a row and column corresponding to the newly formed cluster. The distance between the new cluster, denoted (r,s) and old cluster(k) is defined in this way: d[(k), (r,s)] = min (d[(k),(r)], d[(k),(s)]).
5) If all the data points are in one cluster then stop, else repeat from step 2).

No apriori information about the number of clusters required and also easy to implement and gives best result in some cases.

E. ECC for secure transmission

The Elliptic Curve Cryptosystem (ECC), whose security rests on the discrete logarithm problem over the points on the elliptic curve. The main attraction of ECC over RSA and DSA is that it takes full exponential time. RSA and DSA take sub-exponential time. This means that significantly smaller parameters can be used in ECC than in other systems such as RSA and DSA, but with equivalent levels of security.
In realistic terms, the performance of ECC can be increased by selecting particular underlying finite fields of particular interest and referred to as the elliptic group mod p, where p is a prime number. This is defined as follows. Choose two nonnegative integers, u and v, less than p that satisfy:

$$4u^3+27v^2 \pmod p \neq 0 \quad (1)$$

Then Ep (x, y) symbolises the elliptic group mod p whose elements (x, y) are nonnegative integers less than p which satisfies the condition:

$$Y^2=x^3+ax+b \pmod p \quad (2)$$

Together with the point at infinity O. The elliptic curve discrete logarithm problem can be stated as follows. Fix a prime p and an elliptic curve such that

$$Q= xP \quad (3)$$

Where xP represents the point P on elliptic curve added to itself x number of times. Then the elliptic curve discrete logarithm problem is to determine x given P and Q. It is relatively easy to calculate Q given x and P, but it is very difficult to determine x given Q and P.
D1 and D2 data bases which are distributed across the world. The entire frame can be compared as a three level architecture. Lower

level contains the data bases and OLTP. Middle level contains the architecture of data warehouse. The higher level contains PPDM framework. The advantage of this approach is that each level is independent of the other with respect to the process of execution. The initial transfer of information from different sources to the data warehouse is done by means of ECC. The first phase in this system is to encode the data base records as m to be sent as an x-y point P_m. The data base D1 and D2 are assumed to be sent by sender S1 and S2. The receiver is R1 who maintains data warehouse. Either S1 or S2 will use the following procedure. It is the point P_m that will be encrypted as a cipher text and subsequently decrypted. Note that we cannot simply encode the message as the x or y coordinate of a point, because not all such coordinates are in $E_p(u, v)$. As with the key exchange system, an encryption/decryption system requires a point G and an elliptic group $E_p(u, v)$ as parameters. Each user S1 or S2 selects a private key n_A and generates a public key

$$P_{A=} n_{A X G} \quad (6)$$

To encrypt and send a message P_m to R1, S1 chooses a random positive integer x and produces the cipher text C_m consisting to the pair of points

$$C_{m=\{X G,} P_{m+ X} P_B \ \} \quad (7)$$

Note that S1 has used R1's public key P_B. To decrypt the cipher text, R1 multiplies the first point in the pair by R1's secret key and subtracts the result from the second point:

$$Pm+ xP_B - n_{B(xG)=} P_{m+x(nBG)-(} n_{B(xG)=} P_m \quad (8)$$

S1 has masked the message P_m by adding xP_B to it. Nobody but S1 knows the value of x, so even though P_B is a public key, nobody can remove the mask xP_B. However, S1 also includes a "clue," which is enough to remove the mask if one knows the private key n_B. For an attacker to recover the message, the attacker would have to compute x given G and xG, which is hard. The same way the data base D2 will be sent by S2 to R1 securely.

Functions of ECC

1. Function to find the multiplicative inverse of an integer for a given prime number:-

For this code, we have used the extended Euclid Algorithm whereby the intermediate terms are less than the prime numbers. This prevents the intermediate terms from exceeding the corresponding prime number.

2. Function to generate the points on an Elliptic curve

As there is constant need for a database of the elliptic curve points, a code to scan all Y co-ordinates that satisfy the elliptic curve equation for the given X co-ordinate has been included. Equation of the elliptic curve: $y^2 \bmod p=(x^3+ax+b) \bmod p$

Where ,p is a prime number

Alogorithm:inputs p,a,b

a.enter the input data

b. x=[0: p-1]

c.For each value of x, check which values of y from 0 to (p-1) satisfies the equation.

d.Display the required point.

3. Function to find the public key
4. Function for encoding and decoding

IV. CONCLUSION AND FUTURE WORK

huge amounts of data which is distributed across different locations has lead to an interest in the development of reliable proper security as growing capacity to track and collect over communication, and also data mining algorithms which preserve user privacy. In this paper, we have proposed novel frameworks for privacy preserving, ECC and data distortion technique are used in this process. ECC has been used for secure transmission of data from source to destination which ensures security during transmission and data distortion mechanism for PPDM. This unified approach gives better security and privacy of data as compared with rest of the approaches in literature as most of the earlier research was either on to secure transmission or on to PPDM approach.

We have used data distortion mechanism for PPDM but there are also other approaches in this area, such as random rotation based data perturbation, kernel k-means clustering algorithm, and retention replacement methods, it is interesting to analyze how these methods can be used in the above frame work. Many motivating and imperative directions are worth exploring. This approach can

also be used as a framework for the next researchers on Data Warehouse. Current Warehouse data distributes requires security during transmission and privacy of confidential information. Future research also includes taking actual data and analyzing communication speed and the overhead involved by adding ECC on PPDM. In our framework, Association rule mining is used there is a need to have PPDM mechanism to be designed for other mining algorithms like clustering also. Further research may focus on the combination of randomization and cryptographic methods to get better results in PPDM ensuring high Data Mining results.

REFERENCES

[1] Alani, M.M.," A DES96 - improved DES security ", 7th International Multi-Conference on Systems, Signals and Devices, Amman, 27-30 June 2010.
[2] Seung-Jo Han, Heang-Soo Oh, Jongan Park," IEEE 4th International Symposium on Spread Spectrum Techniques and Application Proceedings ", 22-25 Sep 1996.
[3] Manikandan. G, Rajendiran.P, Chakarapani.K, Krishnan.G, Sundarganesh.G,"A Modified Crypto Scheme for Enhancing Data Security", Journal of Theoretical and Advanced Information Technology, Jan 2012.
[4] Shah Kruti R., Bhavika Gambhava,"New Approach of Data Encryption Standard Algorithm", International Journal of Soft Computing and Engineering (IJSCE), ISSN: 2231-2307, Volume-2, Issue-1, March 2012.
[5] Govind Prasad Arya, Aayushi Nautiyal, Ashish Pant, Shiv Singh, Tishi Handa,"A Cipher Design with Automatic Key Generation using the Combination of Substitution and Transposition Techniques and Basic Arithmetic and Logic Operations",The SIJ Transactions on Computer Science Engineering & its Applications (CSEA), Vol. 1, No. 1, March-April 2013.
[6] Measurement of Cryptographic Primitives on Palm Devices", College of Computer Science, Northeastern University, Boston, MA 02115, USA.
[7] Adi Shamir Ronald Rivest and Len Adleman, "A method for obtaining digital signatures and public-key cryptosystem", Communications of the ACM, 21:120–126, 1998.
[8] Anony mizer.com:http://www.anonymizer.com.
[9] Privacy-preserving Distributed Mining of Association Rules on Horizontally Partitioned Data Murat Kantarcıoglu and Chris Clifton, *Senior Member, IEEE*
[10] Assuring Privacy when Big Brother Murat Kantarcıoglu Chris Clifton
[11] Privacy Preserving Association Rule Mining in Vertically Partitioned Data Jaideep Vaidya & Chris Clifton
[12] Privacy Preserving Data Mining Yehuda Lindell & Benny Pinkas*y*
[13] k-anonymity: Algorithm and Hardness, Gagan Aggarwal, Tomas Feder, Stanford University
[14] Towards Standardization in Privacy Preserving Data Mining, Stanley R. M. Oliveira and Osmar R Zaiane, University of Alberta, Edmonton, Canada