ISSN 2395-1621

Secure Data Mining Using Homomorphic Encryption

[#1]Raunak Joshi,[#2] Bharat Gutal

[1]raunakjoshi007@gmail.com
[2]b.gutal55@gmail.com

[#12]Student Computer Enginering Department, Pune University
PVPIT Pune Maharashtra India

ABSTRACT

Data Privacy is one of the major issues while storing the Data in a database environment. Data Mining based attacks, a major threat to the data, allows an adversary or an unauthorized user to infer valuable and sensitive information by analyzing the results generated from computation performed on the raw data. This presents an approach to mine the data securely using k-means algorithm even in the presence of adversaries. This approach assumes that the data is not stored in a centralized location but is distributed to various hosts. This proposed approach prevents any intermediate data leakage in the process of computation while maintaining the correctness and validity of the data mining process and the end results.

Keywords— Data Mining, Security, K-means, Encryption

ARTICLE INFO

Received : 11th Septmber 2015

Received in revised form :

13th Septmber 2015

Accepted : 14th Septmber 2015

Published online :

18th Septmber 2015

I. INTRODUCTION

Data Mining based attacks, a major threat to the data, allows an adversary or an unauthorized user to infer valuable and sensitive information by analyzing the results generated from computation performed on the raw data.

The homomorphic public key encryption is a cryptographic system that allows the performance of a set of operations on the data when they are encoded, resulting in its data appearing in plain text.The approach is able to maintain the correctness and validity of the existing k-means to generate the final results even in the distributed environment.

II. PROBLEM STATEMENT

This Problem statement proposes a secure k-means data mining approach assuming the data to be distributed among different hosts preserving the privacy of the data. The approach is able to maintain the correctness and validity of the existing k-means generate the final results even in the distributed environment**.** A new approach of modern cryptography, defined as the Homomorphic Encryption allows for the encrypted data to be arbitrarily computed which is a solution that aims to preserve the security, confidentiality and data privacy. This system proposes methods that ensure the confidentiality and privacy in the mining of databases based on fully homomorphic encryption

III. ALGORITHMS

K-means Algorithm for data mining
AES Algorithm for homomorphic encryption

The approach is able to maintain the correctness and validity of the existing k-means to generate the final results even in the distributed environment

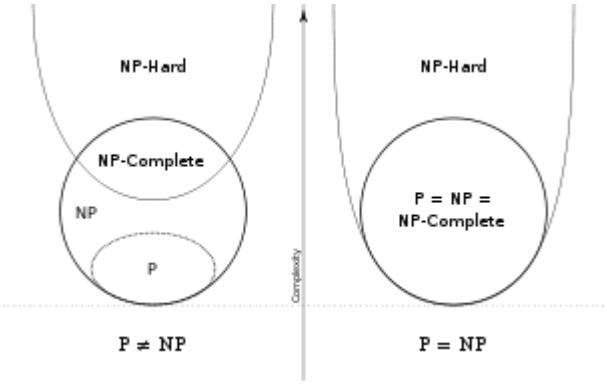
IV. FEASIBILITY ASSESSMENT USING NP-HARD

NP-hard (non-deterministic polynomial-time hard), in computational complexity theory, is a class of problems that are, informally, "at least as hard as the hardest problems in NP". More precisely, a problem H is NP-hard when every problem L in NP can be reduced in polynomial time to H.As a consequence, finding a polynomial algorithm to solve any NP-hard problem would give polynomial algorithms for all the problems in NP, which is unlikely as many of them are considered hard.

A common mistake is thinking that the NP in "NP-hard" stands for "non-polynomial". Although it is widely suspected that there are no polynomial-time algorithms for NP-hard problems, this has never been proven. Moreover, the class NP also contains all problems which can be solved in polynomial time.

A decision problem H is NP-hard when for any problem L in NP, there is a polynomial-time reduction from L to H. An equivalent definition is to require that any problem L in NP can be solved in polynomial time by an oracle machine with an oracle for H. Informally, we can think of an algorithm that can call such an oracle machine as a subroutine for solving H, and solves L in polynomial time, if the subroutine call takes only one step to compute. Another definition is to

require that there is a polynomial-time reduction from an NP-complete problem G to H. As any problem L in NP reduces in polynomial time to G, L reduces in turn to H in polynomial time so this new definition implies the previous one. It does not restrict the class NP-hard to decision problems, for instance it also includes search problems, or optimization problems.



P Class of problems:

Solutions to P class of problems have deterministic algorithms running in polynomial.

NP Class of problems :

- NP = Non-Deterministic polynomial time
- NP means verifiable in polynomial time.

If there is a fast solution to the search version of a problem then the problem is said to be Polynomial-time, or P for short. If there is a fast solution to the verification version of a problem then the problem is said to be Non deterministic Polynomial time or NP for short. Clustering algorithms are generally heuristic in nature and are often polynomial in time.

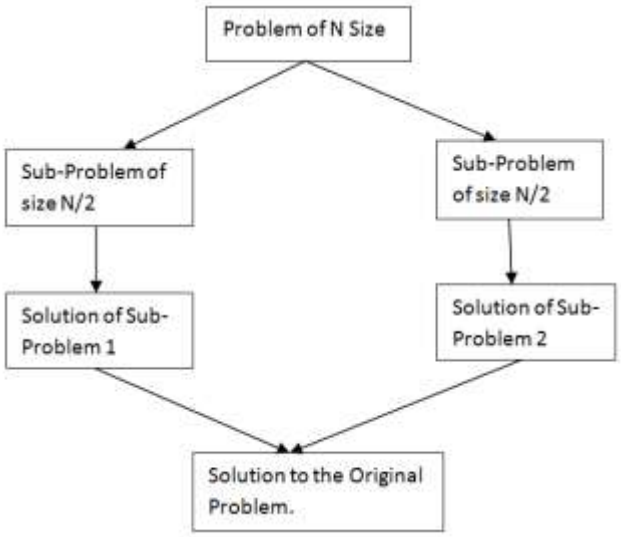
In the k-means problem, we are given a finite set S of points, and integer $k \geqslant 1$, and we want to find k points (centre's) so as to minimize the sum of the square of the Euclidean distance of each point in S to its nearest center. This is done in polynomial time.

There are some complexity-theoretic reasons to believe that cryptography can't be based on NP-completeness. Basically it all boils down to the mismatch between average-case hardness required for cryptography, and worst-case hardness required for NP-hardness. Moreover, many hard algebraic problems which serve as the basis for cryptography are in NP ∩ co-NP . Such problems cannot be NP-complete unless NP = co-NP. Finally, AES is a finite-domain function. It is invertible distinguishable from a random permutation in (large but) constant time. The definitions of P, NP, etc., refer to asymptotic behaviour that is, as the input size grows to infinity. Because of the algebraic structure of AES, it is probably possible to define a "generalized AES" for infinitely many key lengths.

Clustering algorithms are generally heuristic in nature and are often polynomial in time. We are using k means clustering algorithm for data

mining and AES algorithm for homomorphic encryption so our project comes under NP hard Problem.

V. USE OF DIVIDE AND CONQUER STRATEGIES



Divide & conquer is a general algorithm design strategy with a general plan as follows:

Divide: A problem's instance is divided into several smaller instances of the same problem, ideally of about the same size.

Recur: Solve the sub-problem recursively.

Conquer: If necessary, the solutions obtained for the smaller instances are combined to get a solution to the original instance.

Diagram shows the general divide & conquer plan

Advantages of Divide & Conquer technique:

For solving conceptually difficult problems like Tower Of Hanoi, divide & conquer is a powerful tool.

Results in efficient algorithms.

Divide & Conquer algorithms are adapted for execution in multi-processor machines.

Results in algorithms that use memory cache efficiently.

In our project we dividing our database using horizontal partitioning method so naturally there will be a good performance, so we are using one type of divide & conquer strategy.

VI. CONCLUSION

Security and privacy is the major issue concerning the clients as well as of services as a lot of confidential and sensitive data is stored which can provide valuable information to an attacker. This proposes a method to solve the privacy issues of the database. It assumes that the user data is distributed on two hosts and performs a combined k-means clustering using the Homomorphic encryption system for security purpose so as to prevent any interpretation of intermediate results by an attacker. The proposed approach can further be extended by adding a digital signature or hashing technique to authenticate the third party so as to prevent an adversary from posing as the third party to host's. Also it can be generalized or extended to more number of hosts if required.

REFRENCES

1] A. Fox, R. Griffith, A. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, and I. Stoica, "Above the clouds: A Berkeley view of cloud computing." Dept. Electrical Eng. and Comput. Sciences,

University of California, Berkeley, Rep. UCB/EECS 28 (2009): 13.

[2] R. Chow, P. Golle, M. Jakobsson, E. Shi, J. Staddon, R. Masuoka, and J. Molina, "Controlling data in the cloud: outsourcing computation without outsourcing control." In Proceedings of the 2009 ACM workshop on Cloud computing security, pp. 85-90. ACM, 2009.

[3] D. J. Solove, "I've got nothing to hide and other misunderstandings of privacy," San Diego L. Rev. 44 (2007): 745.

[4] P. K. Rexer, "Data miner survey highlights the views of 735 dataminers" 2010.

[5] C. Su, F. Bao, J. Zhou, T. Takagi, and K. Sakurai, "Privacy-preserving two-party k-means clustering via secure approximation." InAdvanced Information Networking and Applications Workshops, 2007,

AINAW'07. 21st International Conference on, vol. 1, pp. 385-391. IEEE, 2007.

[6] Md. Riyazuddin , Dr.V.V.S.S.S.Balaram , Md.Afroze , Md.JaffarSadiq , M.D.Zuber. "An Empirical Study on Privacy Preserving Data Mining". International Journal of Engineering Trends and Technology (IJETT). V3(6):687-693 Nov-Dec 2012. ISSN:2231-5381

[7] K. Che, and L. Liu, "A random rotation perturbation approach to privacy preserving data classification." (2005).

