Semantically Secure Encrypted Relational Data In Cloud Using C4.5 Classification

Pranali D. Desai [1], Prof. V. S. Wadne [2]

[1] PG Student, Department of Computer Engineering
JSPM's Imperial College of Engineering and Research, Pune, India
[2] Asst. Professor, Department of Computer Engineering
JSPM's Imperial College of Engineering and Research, Pune, India

---***---

Abstract - *Data mining is the procedure of finding the information from the large size database. Data mining commonly used as a part of different fields, for example, exploratory examination, banking, medicines and among government workplaces. Classification is one of regular task of the data mining applications. There are a variety of theoretical and practical solutions for classification have been proposed, from last couple of years analyst address different privacy issues. User can outsource the encrypted data to the cloud and also data mining task to the cloud because of increase in Cloud computing. But if the data on cloud is in encrypted form in the existing techniques used for privacy preserving classification are not applicable. So in proposed system, we worked on solving the classification problem over encrypted data. From last several years numerous privacy-preserving category methods have been offered to secure user privacy, we introduced novel algorithm called C 4.5 to solve the issues of classification over encrypted data. In our system to outsource secured information in the cloud we proposed novel algorithm called C 4.5. Proposed algorithm is not only time efficient but also it gives more accurate results.*

Key Words: Security, k-NN classifier, outsourced databases, encryption

1. INTRODUCTION

As of late the data mining has been received a more consideration. Today's digital infrastructure gives another method for putting away, managing out, and dissemination information. Actually, we can store our data on remote servers, access dependable and able service provider, and use computing authority accessible at multiple locations across the network system. The result of a lengthy procedure of assessment and item enhancement is nothing but Data mining system.

In data mining applications classification is one mostly used tasks. It is most common that, company or an organization delegate their data as well as their computational operations to the cloud. without being affected by its advantages that the cloud provide, because of its lack of privacy and security issues companies are avoiding to use of those advantages.

The data requires to be encoded or encrypted before transferring to the cloud, when data are highly exposed. When data are encrypted using any encryption scheme and performing any data mining tasks on that encrypted data becomes very difficult without ever decrypting the data. On a cloud data mining, when the record is a part of a data mining process encrypted data also required to preserve a client's record. Even though cloud can also abstract useful and sensitive information about the data items by examining the data access patterns even if the data are encrypted.

Existing privacy preserving classification techniques are not applicable as the data on cloud is encrypted form that is the main problem. Encryption is the technique of encoding data in such a manner that only authenticated user can access data in cryptography. In an encryption scheme, using an encryption algorithm encoded data or message is referred to as plaintext, produce cipher text that can only be read if decrypted.

2. RELATED WORK

In paper [2]. Authors introduced a practical ignorant data access protocol with accuracy. The key lie in new constructions and well cultured reshuffling protocols that give in practical computational complexity (to O (log n log log n)) and storage expenditure (to O (n)). author also bring in a practical implementation that allows a throughput of numerous queries per second on 1Tbyte+ databases, with full computational privacy and accuracy, orders of magnitude faster than present approaches.

In paper [3], author elaborate the working of a different of Gentry's fully homomorphic encryption scheme implemented the underlying "somewhat homomorphic" scheme, but were not able to execute the bootstrapping functionality that is needed to get the complete scheme to work. There are number of optimizations that allow us to implement all aspects of the scheme, including the bootstrapping functionality.

In paper [4], author proposes a scheme that allows one to calculate circuits over encrypted data without being able to decrypt. Result comes in three steps. First, to build an encryption scheme that permits evaluation of arbitrary circuits, a general result that, it suffices to construct an encryption scheme that can evaluate its own decryption circuit; call a scheme that can evaluate its (augmented) decryption circuit boots trappable.

In paper [5], for developing privacy preserving applications author has proposed a novel approach. The SHAREMIND system introduces numerous ideas for improving the efficiency of both the applications and their growth process depends on secure multi-party computation, instead of the standard finite field computation protocols works over elements in the ring of 32-bit integers that is the main contribution of the system. Because of this we can build simple and efficient protocols. Due to SHAREMIND's easy to use application development interface programmer can concentrate on the execution of data mining algorithms without worrying about the privacy issues.

In paper [6], authors address the problem of privacy preserving data mining, On the union of two databases without enlightening any unnecessary information in which two parties owning confidential databases wish to run a data mining algorithm. Our work is motivated by the need to both defend confidential information and enable its use for research or other purposes. The problem stated above is a example of secure multi-party computation; this can be solved using known generic protocols. Still, data mining algorithms are complex and, the input usually having large data sets.

In paper [7], the problem of secure distributed classification is an significant one. In many cases data is divided between multiple organizations. While hiding their training data these organizations may want to utilize all of the data to create more accurate predictive models The Naive Bayes Classifier is a simple but efficient baseline classifier.

In paper [8], author present a framework for mining association rules from transactions consisting of categorical items where the data has been randomized to preserve privacy of individual transactions. While it is reasonable to recover association rules and preserve privacy using a straight forward randomization, the rules that are found can be used to find privacy leaks.

In paper [9], allocated privacy preserving data mining methods are critical for mining a number of databases with a lowest information revelation. We present a formation along with a general model as well as multi-round algorithms for exploration side to side partitioned databases using a comfort protecting k Nearest Neighbor (kNN) classifier.

In this paper [10], auther talk about the protected computations on a secured databases and suggest a SCONEDB (Secure Computation ON an Encrypted Database) design, which catches the effectiveness and security specifications. It focus on the issue of k-nearest neighbor (kNN) calculations on secured databases. We create a new asymmetric scalar-product-preserving encryption (ASPE) that maintains a special type of scalar item. We use APSE to create two protected techniques those assistance kNN calculations on secured data; each of these techniques is proven to avoid realistic strikes of a different qualifications knowledge level, at a different expense cost. Comprehensive efficiency research is performed to assess the expense and the efficiency of the techniques.

3. IMPLEMENTATION DETAILS

3.1. System Architecture

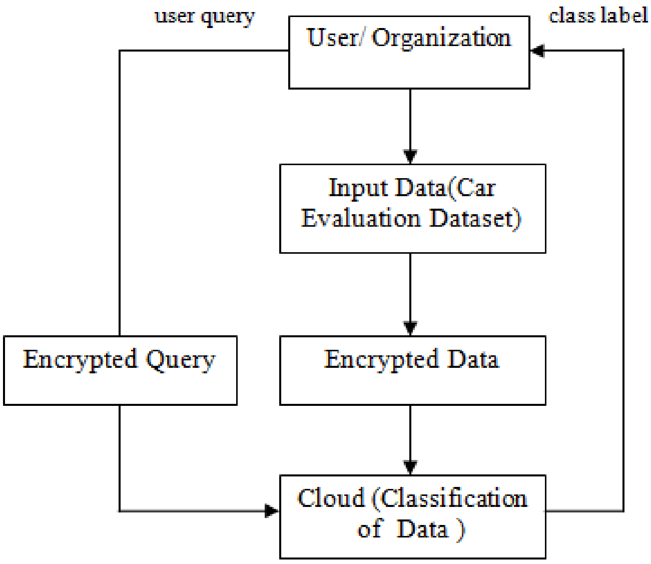

Figure 1: System Architecture

We introduced novel algorithm called C 4.5 to solve the problem of classification over encrypted data. We projected methods to successfully solve the DMED difficulty assuming that the encrypted data are outsourced to a cloud. Our focus is on the classification difficulty since it is one of the most common data mining tasks. Because each classification technique has their own advantage, to be concrete, we concentrated on implementing the k-nearest neighbor classification and C4.5 classification method over encrypted data in the cloud computing environment. The proposed system protects the confidentiality of data, privacy of user's input query, and hides the data access patterns.

3.2. Algorithm:

- C4.5 Algorithm

C4.5 is an classification algorithm used to generate a decision tree developed by Ross Quinlan. It is an extension of ID3 algorithm that accounts for unavailable values, continuous attribute value ranges, and pruning of decision trees, rule derivation, and so on. It is also use as a statistical classifier. In non-binary case the ID3 algorithm favors the attributes with huge area of possible values.

To deal with this problem C4.5 algorithm is used. In the C4.5 algorithm the proportion of the information gain and the split information is projected as the split measure function.

There are various steps of C4.5 algorithm :

1) first we have to test for base cases

2) Then for each attribute x, we discover the normalized information gain ratio from splitting on a

3) Let x_best be the attribute with the maximum normalized information gain.

4) Then build a decision node that split on x_best.

5) Next return on the sub lists generated by splitting on x_best, and insert those nodes as children of node.

3.3. Mathematical model

$$M = (Q, \sum, P, q_0, F)$$

Where,
Q is the set of States
$\sum$ Is the set of inputs
P State Transition Table
q_0 is the initial State
F is the final State

- Q: { S_0, S_1, S_2}

Where,
S_0: Encryption
S_1: Classification in cloud
S_2: Class labels

- { Ed, Eq, Cf }

Where,
Ed: Encrypted data
Eq: Encrypted query
Cf: C4.5 Classifier

- q_0 : {S_0}
- F: {S_2}

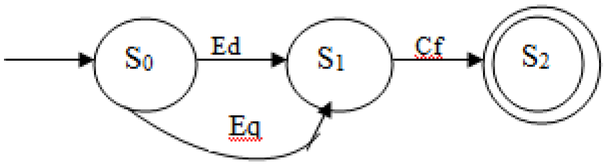

Figure.2: State Diagram

4. RESULTS

Below graph shows the Accuracy comparison. Figure 3 graph demonstrates the Accuracy comparison graph. This graph proves that the proposed algorithm is more accurate as compared with the previous algorithm.

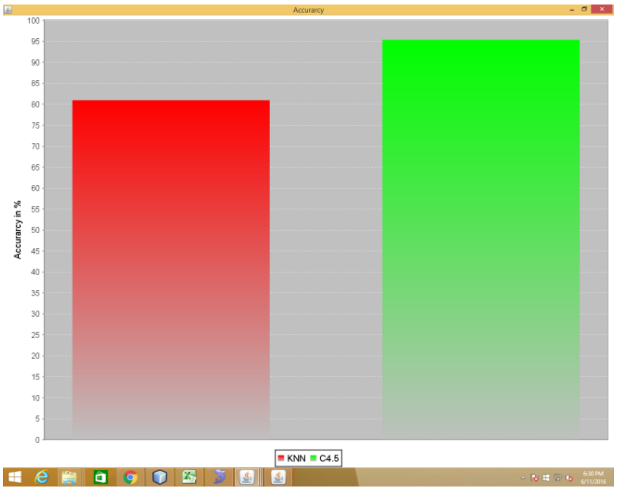

Figure 3: Accuracy Comparison

Below graph shows the Time comparison. This graph represents Time variation between proposed and existing algorithm. Proposed system improves the time efficiency as compared to existing system as shown in the below graph.

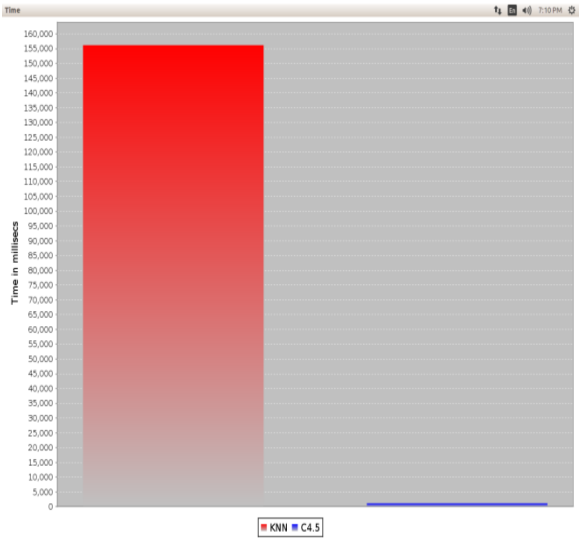

Figure 4: Time Comparison

5. CONCLUSION

As increased popularity of Cloud computing, user can outsource the encrypted data to the cloud and perform data mining tasks to the cloud. The difficulty with existing privacy preserving classification techniques is that they are not appropriate as the data on the cloud is in encrypted structure.

To solve the problem of classification over encrypted data we proposed algorithm called C 4.5 It is used for outsource secured information in the cloud. C 4.5 algorithm is time efficient as well as it gives more accurate results.

ACKNOWLEDGEMENT

I express my gratitude towards Prof. V. S. Wadne , project guide and Prof. R. N. Phursule , P.G. coordinator and Prof. S. R. Todmal, Head of Department of Computer Engineering , Imperial College Of Engineering And Research Wagholi, Pune who guided and encouraged me in completing this paper. I would also like to thanks our Principal Dr. S. V. Admane

REFERENCES

[1] Bharath K. Samanthula, , Yousef Elmehdwi, and Wei Jiang, " k-Nearest Neighbor Classification over Semantically Secure Encrypted Relational Data " IEEE Transactions On Knowledge And Data Engineering, Vol. 27, No. 5, May 2015

[2] P. Williams, R. Sion, and B. Carbunar, "Building castles out of mud: Practical access pattern privacy and correctness on untrusted storage," in Proc. 15th ACM Conf. Comput. Commun. Security, 2008, pp. 139-148.

[3] C. Gentry, "Fully homomorphic encryption using ideal lattices," in Proc. 41st Annu. ACM Sympos. Theory Comput., 2009, pp. 169-178.

[4] C. Gentry and S. Halevi, "Implementing gentry's fully-homomorphic encryption scheme," in Proc. 30th Annu. Int. Conf. Theory Appl. Cryptographic Techn.: Adv. Cryptol., 2011, pp. 129-148.

[5] D. Bogdanov, S. Laur, and J. Willemson, "Sharemind: A framework for fast privacy preserving computations," in Proc. 13th Eur. Symp. Res. Comput. Security: Comput. Security,2008, pp. 192-206.

[6] R. Agrawal and R. Srikant, "Privacy-preserving data mining," ACM Sigmod Rec., vol. 29, pp. 439-450, 2000.

[7] P. Zhang, Y. Tong, S. Tang, and D. Yang, "Privacy preserving Naive Bayes classification," in Proc. 1st Int. Conf. Adv. Data Mining Appl., 2005, pp. 744-752.

[8] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, "Privacy preserving mining of association rules," Inf. Syst., vol. 29, no. 4, pp. 343-364, 2s004.

[9] L. Xiong, S. Chitti, and L. Liu, "K nearest neighbor classification across multiple private databases," in Proc. 15th ACM Int. Conf. Inform. Know. Manage. 2006, pp. 840–841.

[10] W. K. Wong, D. W.-l. Cheung, B. Kao, and N. Mamoulis, "Secure kNN computation on encrypted databases," in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2009, pp. 139–152.