International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 - 6435 (Online) Volume 2, Issue 4, April 2014)

To Maintain Privacy Using Add Multiply and K-Mediod

Isha Sahni[1], Prof. Hare Ram Shah[2]
[1,2]*Department of Computer Science and Engineering, Gyan Ganga Institute of Technology & Sciences, Jabalpur (M.P.)*

***Abstract*— To conduct data mining, there is often need to collect data from various parties. Privacy concerns may prevent the parties from directly sharing the data and some types of information about the data. Data collection is a necessary step in data mining process. Due to privacy reasons, collecting data from different parties becomes difficult. The challenge behind this is that multiple parties should collaboratively conduct data mining without breaching data privacy. Objective of PPDM is to preserve confidential information by applying data mining tasks without modifying the original data. In this paper the study of privacy preserving using Add Multiply protocol and K-Mediod, based on homomorphic encryption and digital envelope techniques, defined to exchange the data while keeping it private. A secure protocol for multiple parties to conduct the desired computation is developed. The solution is distributed, i.e., there is no central, trusted party having access to all the data.**

***Keywords*-- PPDM, Decision Tree Algorithm, K-Mediod Clustering, Add To Multiply Protocol Based Homomorphic Encryption, Secure Multi-Party Computation**

I. INTRODUCTION

Data mining is the procedure of extracting hidden information from large data sets. Recent advances in data collection, data dissemination and related technologies have inaugurated a new era of research where existing data mining algorithms should be reconsidered from the point of view of privacy preservation. As a data mining function, clustering is the process of grouping a set of physical and abstract objects into classes of similar objects. The advantage of using clustering based process are that it is adaptable to changes and helps out useful features that distinguish different groups. Many numbers of privacy-preserving data mining techniques have newly been projected which take either a cryptographic [9] or a statistical approach. Secure multi-party computation is used in the cryptographic approach which ensures strong privacy and accuracy. But, this approach typically suffers from its poor performance. The statistical approach has been used to extract the facts from association rules, clustering and decision trees. This approach is very popular because of its high performance. Privacy has become an important issue in Data Mining.

In order to protect the privacy information, the objective of privacy preserving data mining is to hide certain sensitive information so that they cannot be discovered through data mining techniques. The authors deal with the problem of association rule mining which preserve the confidentiality of each database. In order to avoid the privacy, information is broadcasted or been illegally used. Privacy preserving data mining (PPDM) has emerged to address this issue. Most of the techniques for PPDM uses modified version of standard data mining algorithms, somewhere the modification are made using well known cryptographic techniques ensure the required privacy for the application for which the technique was designed. In most cases, the constraints for PPDM are preserving accuracy of the data and the generated models and the performance of the mining process while maintaining the privacy constraints. The numerous procedures used by PPDM [3] can be summarized as below:

1. The data is changed before delivering it to the data miner.
2. The data is circulated between two or more locations which work together using a semi-honest protocol to learn global data mining results without revealing any information about the data at their individual sites.
3. While using a representation to classify data, the classification outcomes are only exposed to the selected party, who does not learn something else, further the classification results, but can check for existence of certain rules without revealing the rules.

II. RELATED WORK

Secure Multi-party Computations (SMC) [3] deal with computing any function on any input in a distributed network. Each participant holds one of the inputs while ensuring that no more information is revealed to a participant in the computation than can be inferred from that participant's input and output. It has been proved that for any polynomial function, there is a secure multi-party computation solution. The approach [4] used is as follows: the function F to be computed is first represented as a combinatorial circuit, and then the parties run a short protocol for every gate in the circuit.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 - 6435 (Online) Volume 2, Issue 4, April 2014)

Every participant gets corresponding shares of the input wires and the output wires for every gate. This approach, though appealing in its generality and simplicity, is highly impractical for large data sets. Following the idea of secure multiparty computation, people designed privacy-oriented protocols for the problem of privacy- preserving collaborative data mining.

From the privacy protection point of view, the k-mediod clustering seems more challenging than the k-means clustering because we always use the real instance to compute the distances in k-mediod clustering while we use the *mean instance* whose values are the means of the real instance values to compute the distances.

In PPDM, there are two methods to protect actual data from being disclosed, i.e. data perturbation methods (randomization-based techniques), the secure computation method (encryption technique). Data perturbation [1, 3, 7] techniques are used to protect individual privacy for classification, by adding random values from a normal distribution of mean 0 to the actual data value. One problem with this approach is the existing tradeoff between the privacy and the accuracy of the results. While secure computation has an advantage over perturbation in that it provides accurate results and not approximation, it requires considerable computation and communication overhead for each secure computation step.

In general information security model [2], the threats and security fears come from inside attackers and outside attackers. In data mining, the inside attackers are the collaborative parties and the outside attackers are the other network attackers. Prevention of inside attackers is different from prevention of outside attackers in that the inside attackers usually have more knowledge about private data than outside attackers.

The research shows that by protecting the actual data from being disclosed, one approach is to alter the data in a way that actual individual data values cannot be recovered, while certain computations can still be applied to the data. Due to the fact that the actual data are not provided for the mining, the privacy of data is preserved. This is the core idea of randomization-based techniques. The random perturbation technique is usually realized by adding noise or uncertainty to actual data such that the actual values are prevented from being discovered. Since the data no longer contains the actual values, it cannot be misused to violate individual privacy.

The privacy preservation means that multiple parties collaboratively get valid data mining results while disclosing no privacy data to each other or any party who is not involved in the collaborative computations.

To measure the performance of the system i.e. by reducing the cost computation and time using Decision tree [3] for both realized dataset and unrealized dataset.

III. Notion Of Privacy

The need for privacy is sometimes due to law (e.g., for medical databases) or can be motivated by business interests. However, there are situations where the sharing of data can lead to mutual benefit. Despite the potential gain, this is often not possible due to the confidentiality issues which arise. It is well documented that the unlimited explosion of new information through the Internet and other media has reached a point where threats against privacy are very common and deserve serious thinking. Consider a scenario that there are several hospitals involved in a multi-site medical study. Each hospital has its own data set containing patient records. These hospitals would like to conduct data mining over the data sets from all of hospitals with the goal of more valuable information that would be obtained via mining the joint data set. Due to privacy laws, one hospital cannot disclose their patient records to other hospitals.

Data mining (DM) is used in the investigation of activities by using mining techniques. It is widely used by researchers for business and science applications. Because the Data composed from individuals are key essential for making decision or recognition based applications

In practice, there are many environments where privacy preserving collaborative data mining is desirable. For example, several pharmaceutical companies have invested a significant amount of money in conducting genetic experiments with the goal of discovering meaningful patterns among genes. To increase the size of the population under study and to reduce the cost, companies decide to collaboratively mine their data without disclosing their actual data because they are only interested in limited collaboration; by disclosing the actual data, a company essentially enables other parties to make discoveries that the company does not want to share with others. In another field, the success of homeland security aiming to counter terrorism depends on a combination of strength across different mission areas, effective international collaboration and information sharing to support a coalition in which different organizations and nations must share some, but not all, information. Information privacy thus becomes extremely important and our technique can be applied. In the Internet era, collaborative data mining is becoming a popular way to extract useful knowledge from large databases.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 - 6435 (Online) Volume 2, Issue 4, April 2014)

Because of the establishment of privacy laws and privacy concerns of individuals, collaborative data mining cannot be achieved without using privacy protection technologies.

IV. TECNIQUES

In this section, we first describe the cryptographic encryption.

A. *Encryption*

Encryption is a well-known technique for preserving the confidentiality of sensitive information. In comparison with the other techniques described, a strong encryption [4,9] scheme can be more effective in protecting the data privacy. In this technique, add to multiply protocol is applied based on homomorphic encryption and digital envelope techniques to privacy-maintaining data mining.

1) *Homomorphic Encryption:*

A cryptosystem is homomorphic with respect to some operation * on the message space if there is a corresponding operation *′ on the cipher text space such that e(m) *′ e(m′) = e(m * m′). In privacy-oriented protocols, the use of additive homomorphism [4], a new mechanism based on the idea that it is hard to factor number n = pq where p and q are two large prime numbers. The proposed encryption scheme when compared with existing public-key cryptosystems. The results show that the encryption process is comparable with the encryption process of RSA in terms of the computation cost; the decryption process is faster than the decryption process of RSA.

In this technique [1, 5, 6, 7], the following property of the homomorphic encryption functions: e (m1) × e(m2) = e(m1 + m2) where m1 and m2 are the data to be encrypted. Because of the property of associativity, e(m1 + m2 + · · +mn) can be computed as:

e(m1) × e(m2) × ·· ·×e(mn) where e(mi) ≠ 0.

That is d(e(m1 + m2 + ·· ·+mn)) = d(e(m1) × e(m2) × ·· ·×e(mn)) ----------(1)

Note: d (e (m1) m2) = d (e (m1 × m2)) ---------- (2) where × denotes multiplication.

2) *Digital Envelope:*

A digital envelope [1, 5] is a random number (or a set of random numbers) only known by the owner of private data.

To hide the private data in a digital envelope [4, 6, 7], conduct a set of mathematical operations between a random number (or a set of random numbers) and the private data. The mathematical operations could be addition, subtraction, multiplication, etc.

For example, assume the private data value is $\tilde{o}$. There is a random number R which is only known by the owner of $\tilde{o}$. The owner can hide $\tilde{o}$ by adding this random number, e.g., $\tilde{o} + R$.

B. *K-Mediod Clustering:*

Clustering is the process of grouping a set of objects into classes or clusters so that objects within a cluster have similarity in comparison to one another, but are dissimilar to objects in other clusters. In other words, clustering is a process of finding natural groupings in a set of data. There are many clustering [8] algorithms such as k-means method, K-Mediod method, etc. The focus is on k-mediod method since it allows arbitrary objects that are not limited to numerical attributes. In k-mediod clustering, a cluster is denoted by one of its points. It is an easy solution in that it covers any attributes type and that mediod are resistant against outliers. Once Mediod are chosen, clusters are defined as subsets of points close to respective Mediod, and the objective function is described as the distance between a point and its mediod.

The purposes of both algorithms are different. The purpose of k-Mediod clustering is to compute the distance between one instance and each of k mediod, then assign this instance to a mediod with the smallest distance value, and repeat this process for all the instances. However, the purpose of k-nearest neighbor classification is to compute the distance between a query instance and each instance in the dataset, select k closest instances, and assign a class label for the query instance by the majority class principle.

1) *Privacy-Preserving K-Nearest Neighbor Classification:*

The k-nearest neighbor classification [1] is an instance based learning algorithm that has been shown to be very effective for a variety of problem domains. The objective of k-nearest neighbor classification is to discover k nearest neighbors for a given instance, then assign a class label to the given instance according to the majority class of the k nearest neighbors [2]. The nearest neighbors of an instance are defined in terms of a distance function such as the standard Euclidean distance. Let an arbitrary instance x be described by the feature vector $<a1(x), a2(x),\ldots, ar(x)>$, where $ai(x)$ denotes the value of the ith attribute of instance x. Then the distance between two instances xi and xj is defined as $dist\ (xi; xj)$, where

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 - 6435 (Online) Volume 2, Issue 4, April 2014)

$$dist(x_i, x_j) = \sqrt{\sum_{q=1}^{r} (a_q(x_i) - a_q(x_j))^2}.$$

C. Privacy-Preserving Add To Multiply Protocol (Ppatmp):

Input: The private number [4] for both the sender and recipient is considered as the input. Suppose senders private no. is x and that of receivers is y.

Output: The secret output for both the sender and recipient is considered to be u and v, such that $x+y=u \cdot v$.

PPAtMP based on Homomorphic Encryption System Homomorphic Encryption system allows computing the sum of encrypted data without decrypting them.

The main purpose of using ads to multiply protocol is to secure the message during transmission against forgery.

V. Unrealized Training Set And Decision Tree Learning Algorithm

Different decision trees [3] can be build from the same training set, because of the undetermined selection criteria of the test attribute in the recursive case. The efficiency of a test element or attribute can be determined by its classification of the training set. A perfect attribute splits the outcomes as an exact classification, which achieves the goal of decision-tree learning [6]. Diverse criteria are used to select the "best" attributes, e.g. Gini index. Among these criteria, information gain is commonly used for measuring distribution of random events. Iterative Dichotomiser3 (ID3) selects the test attribute based on the information gain provided by the test outcome. Information gain measures the change of uncertainty level after a classification from an attribute. Fundamentally, this measurement is rooted in information theory.

Input: Set of training samples (Ts): R1, R2, …, Rm and set of attributes a1, a2, ….., am

Default: default value for target predicate

Output: Decision Tree

Procedure build-tree (Ts, attribute, default)

1. If Ts is empty then return default
2. Default<-Majority-Value Ts
3. If Hm(Ts) then return default
4. Else if attribute is empty then return default
5. Else
6. Best choose-attribute(attribute,Ts)
7. Tree a new decision tree with root attribute best
8. For each value Vi of best do
9. Ti dataset in Ts as best = Ki
10. Subtree<-Generate-Tree(attribute best, Ts, default)
11. Connect tree and Subtree with a branch labeled Ki
12. Return tree

To unrealized the samples, initialization of both set of input sample dataset and perturbing dataset as empty sets, i.e. Unrealized training set is called. Consistent with the procedure described above, universal dataset is added as a parameter of the function because reusing pre-computed universal dataset is more efficient than recalculating universal dataset. The recursive function unrealized training-set takes one dataset in input sample dataset in a recursion without any special requirement; it then updates perturbing dataset and set of output training data sets correspondent with the next recursion. Therefore, it is obvious that the unrealized training set process can be executed at any point during the sample collection process.

Input: Unrealized training data set

Output: Modified decision tree

If unrealized data set is empty then return default

Default<-Minority-Value

Else

Tree<-best highest value of information gain (root)

Subtree <- tree (root, best size)

Connect tree and Subtree

Return tree

End

Similar to the traditional ID3 [3] algorithm Choose Attribute selects the test attribute using the ID3 criterion based on the information entropies, i.e., select the attribute with the greatest information gain. Algorithm Minority-Value retrieves the least frequent value of the decision attribute, which performs the same function as algorithm Majority-Value of the tradition ID3 approach that is, getting the majority frequent value of the decision attribute of Ts. The decision attribute should be arbitrarily chosen and generate the decision tree by calling the function Generate-Tree.

Attributes: Set of attributes

Default: Default value for target predicate

Output: Tree, Decision Tree

1. If (T' , T^p) is empty then return default
2. Default minority value (T' , T^p)

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 - 6435 (Online) Volume 2, Issue 4, April 2014)

3. If then return default
4. Else if attribute is empty then return default
5. Else
6. Best<-choose-attribute'(attribute, size (T^p))
7. A new decision tree with root attribute best
8. Size<-size/number of possible values Ki in best
9. For each value Vi of best do
10. Ti' dataset in T' as best = Ki
11. T^p' dataset in T^p as best = Ki
12. Subtree Generate-Tree(size, T', T^p, attribute-best, default)
13. Connect tree and Subtree with a branch labeled Ki
14. Return tree

VI. CONCLUSION

Privacy-preserving data mining has generated much research. In this paper, we propose a formal definition of privacy and use homomorphic encryption and digital envelope technique to achieve collaborative data mining without sharing the private data among the collaborative parties and then we present a new scheme to compute the k-nearest neighbor search based on the homomorphic encryption.

The output of the existing system consist of tree and decision tree, so it takes a lot of time to compute a tree first and then a decision tree. Therefore by considering a final decision tree as our output will save the time and computation of cost as compared to the one used in earlier case.

REFERENCES

[1] "A Crypto-Based Approach to Privacy-Preserving Collaborative Data Mining", Justin Zhan and Stan Matwin on February 22,2010.

[2] Jianming Zhu, **"**A New Scheme to Privacy-Preserving Collaborative Data Mining**",** 2009 Fifth International Conference on Information Assurance and Security.

[3] Ms.S.Nithya, Mrs. P.Senthil Vadivu, Volume 2, No.6, June 2013, "Efficient Decision Tree Based Privacy Preserving Approach For Unrealized Data Sets".

[4] R.Raju, R.Komalavalli, V.Kesavkumar, Second International Conference on Emerging Trends in Engineering and Technology, ICETET-09, **"**Privacy Maintenance Collaborative Data Mining –A Practical Approach**".**

[5] "Using Homomorphic Encryption and Digital Envelope Techniques for Privacy Preserving Collaborative Sequential Pattern Mining", Justin Zhan, Carnegie Mellon University on February 23,2010.

[6] **"**Using Homomorphic Encryption For Privacy-Preserving Collaborative Decision Tree Classification**",** Proceedings of the 2007 IEEE Symposium on Computational Intelligence and Data Mining (CIDM 2007), Justin Zhan.

[7] **"**Privacy-Preserving Collaborative Data Mining**"**, Justin Zhan, Carnegie Mellon University, USA, February 23,2010

[8] "Privacy Preserving K-Mediod Clustering", Justin Zhan.

[9] M. Shaneck and Y. Kim, **"**Efficient Cryptographic Primitives for Private Data Mining**",** Proc. 43rd Hawaii Int'l Conf. System Sciences (HICSS), pp. 1-9, 2010.