www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3 Issue 5 May, 2014 Page No. 5594-5599

DISCLOSURE PROTECTION OF SENSITIVE ATTRIBUTES IN COLLABORATIVE DATA MINING

*V. Uma Rani *1, Dr. M. Sreenivasa Rao *2, V. Theresa Vinayasheela *3*

1. Asst. Prof of CSE, School of IT, JNTUH, Hyderabad (A.P), India.
2. Professor of CSE, School of IT, JNTUH, Hyderabad (A.P), India.
3. Lecturer, Dept of Computer Science & Engg., Loyola Academy Degree and PG College, Old Alwal, Secunderabad-10,A.P, India, Part-time M.Tech(CS) student of SIT, JNTUH.

vtheresavinayasheela@gmail.com

ABSTRACT

In collaborative data mining, data sets from various parties are submitted to a third party where they are combined, privacy of each data sets' sensitive attributes are protected and data mining is carried out. In privacy preserving data mining, there is a need to extract knowledge from databases without disclosing information about individuals. Each participant will have sensitive and non-sensitive data in their local database. Therefore the most important challenge in privacy preserving multi party collaborative data mining is how these multiple parties conduct data mining without disclosing each participant's sensitive data. In this paper we propose a two-level encryption algorithm for protecting sensitive attributes from disclosure and a generalization algorithm. This approach guarantees high level privacy with less amount of complexity as compared to the existing methods and also proves to be fast and efficient over dynamic queries.

INTRODUCTION

In privacy preserving data mining, there is a need to extract knowledge from databases without disclosing information about individuals. When common users are involved in data mining all of them transfer their data to a trusted common centre to conduct mining. But it is very difficult for a particular user to trust other users. Hence privacy of each user is of great concern. This process is called Privacy Preserving Collaborative Data Mining. The techniques for this are based on information hiding and data mining. The attributes which are designated as sensitive and not to be disclosed are encrypted by an efficient encryption algorithm by the user with the help of a key which is shared by the third party to enable perform data mining and yet preserve privacy of sensitive attributes. Along with providing encryption facility this project also implements generalization on certain data values upon merging different data sets.

The medical data of patients' from different hospitals can be combined, medical data and health insurance data can be combined etc, where the data sets are integrated based on a common attribute. Once this is achieved, sensitive attributes are protected and data mining is conducted. Consider the hospitals data where the ailment and the treatment amount can be extracted with some interesting information. Similarly consider employee data where different ranges of pay scale can be extracted post protecting the sensitive attributes such as personal details.

1. RELATED WORK

A number of methods have been proposed to address the problem of privacy preserving in collaborative data mining. A detailed analysis of all some of the currently used methods can be found in [1].

The goal of association rule mining is to discover meaningful association rules among the attributes of a large quantity of data. A secure collaborative association rule mining protocol is developed based on homomorphic encryption scheme. In this protocol, the parties do not send all their data to a central, trusted party. Instead, we use the homomorphic encryption techniques to conduct the computations across the parties without compromising their data privacy. However this does not show the privacy level achieved. [2].

Public key distribution scheme introduced by Diffie and Hellman to encrypt and decrypt messages. The security of this system is equivalent to that of the distribution scheme. It also introduces a new digital signature scheme that depends on the difficulty of computing discrete logarithms over finite fields. However it is not yet proved that breaking the system is equivalent to computing discrete logarithms. The public key system can be easily extended to any GF (p^m), but recent progress in computing discrete logarithms over GF (p”) where m is large makes the key size required very large for the system to be secure. Hence, it seems that it is better to use GF(p”) with m = 3 or 4 for implementing a cryptographic system. For the same security level, the size of the public key file and the size of the cipher text will be double the size of those for the RSA system. [3].

2. BACKGROUND

In this section, background details of the methodologies presented earlier are discussed which form the basis for our paper.

3.1 ALGORITHMS

A. Association Rule Mining Algorithms

This paper [2] considered the problem of privacy-preserving through association rule mining. In particular the study on how multiple parties can collaboratively conduct association rule mining on their joint private data has been conducted. . The goal of association rule mining is to discover meaningful association rules among the attributes of a large quantity of data. A secure collaborative association rule mining protocol is developed based on homomorphic encryption scheme. In this protocol, the parties do not send all their data to a central, trusted party. Instead, we use the homomorphic encryption techniques to conduct the computations across the parties without compromising their data privacy. However this does not show the privacy level achieved.

B. Signature Scheme Based on Discrete Logarithms

This paper [3] presents systems that rely on the difficulty of computing logarithms over finite fields. It implements the public key distribution scheme introduced by Diffie and Hellman to encrypt and decrypt messages. The security of this system is equivalent to that of the distribution scheme. It also introduces a new digital signature scheme that depends on the difficulty of computing discrete logarithms over finite fields. However it is not yet proved that breaking the system is equivalent to computing discrete logarithms. The public key system can be easily extended to any GF (p^m), but recent progress in computing discrete logarithms over GF (p") where m is large makes the key size required very large for the system to be secure. Hence, it seems that it is better to use GF(p") with m = 3 or 4 for implementing a cryptographic system. For the same security level, the size of the public key file and the size of the cipher text will be double the size of those for the RSA system.

C. A Crypto-Based Approach

This paper [4] presents a solution for K-nearest neighbor classification with vertical collaboration. The efficiency analysis shows the performance scaling up with various factors such as the number of parties involved in the computation, the encryption key size, the size of data set, etc. This paper, proposes 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. The drawback of this however is that it does not support horizontal collaboration k-nearest neighbor classification.

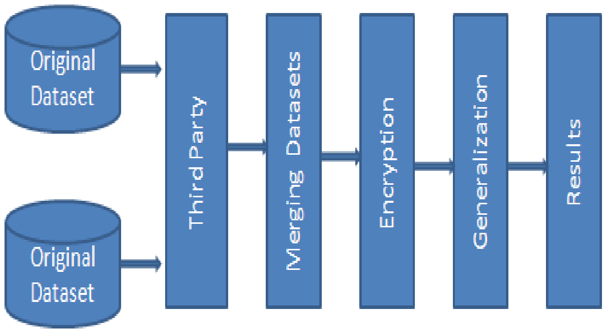
D. A Probabilistic Encryption

This paper presents the idea to find an encryption function E which is easy to compute but difficult to invert unless some secret information, the trapdoor is known. Such a function is called a trapdoor function. To encrypt a message m, anyone can simply evaluate E(m), but only those who knows the trapdoor information can compute m from E(m).

4. OUR NEW APPROACH

Our work introduces two algorithms

First, *a two-level Encryption algorithm* that aims at high level of privacy preserving of sensitive attributes over dynamic queries. XOR encryption algorithm is used that operates on the alphabet set. Second, *Generalization algorithm* which works on minimum and maximum values of the given attribute in the given range to generalize the data.



4.1Encryption Algorithm

The encryption algorithm proposed in this paper works at two levels. First, the key generated will be computed upon by a polynomial function. Second, a bit-wise XOR is performed on the key (which is now transformed as a result of the polynomial function) and the protected column whose privacy is to be preserved.

4.1.1 Key Transformation

The key shared by the participating party in multi-party collaborative data mining is first transformed by applying an encoding methodology according to the following algorithm. The key is now made secure and even if there may be some insiders aware of the key they will never be able to again unauthorized access to the merged data.

- An array S is initialized with the values ranging from 0 to 255.

```
For i = 0 To 255
s(i) = i
Next
```

- A second array K is initialized with the ASCII values of the letters in the key repeatedly for the same length as the array S.

```
 j = 1
For i = 0 To 255
If j > Len(key) Then j = 1
K(i) = Asc(Mid(key, j, 1))
j = j + 1
Next i
```

- The values in the array S are interchanged among themselves according to the steps given below. And as a result the values in the array S are changed.

```
j = 0
For i = 0 To 255
j = (j + s(i) + K(i)) Mod 256
temp = s(i)
s(i) = s(j)
s(j) = temp
Next i
```

- Finally an encrypted value of the key is obtained according to the below given steps.

```
i = 0
j = 0
For x = 1 To Len(key)
j = j + s(i)
Next
rval = j Mod 256
encrypt = rval
```

4.1.2 Encryption using XOR

- The value obtained for the given key is now XORed with each letter of the value of the attribute under consideration.

4.1.3 Decryption using XOR

- The encrypted values are decrypted by following the same procedure of XORing each letter of the values of the attribute under consideration however the encrypt value to decrypt with is obtained using the same key transformation procedure discussed earlier.

4.2 Generalization Algorithm

- This paper proposes a generalization method using minimum and maximum values from the given range of the attribute under consideration. Firstly the algorithm requires to find the minimum and maximum, which is done as follows:

```
int mn = 0; int mx = 0;
int rn = 1;
if (ii == 0)
{
mn = cv;
mx = cv;
}
else
{
if (cv < mn)
mn = cv;
if (cv > mx)
mx = cv;
}
ii++;
```

- Once this is achieved, the next step is to compute a new value in place of the original value in an attempt to not to disclose the original values to the other users post collaborating the datasets. This can be done as follows:

```
int d1 = ((mn * mx) – Item (original)
```

- In the same way the original values can be obtained using the same step however this time the value will be the changed one in order to retrieve the original.

```
int d1 = ((mn * mx) – Item (changed)
```

5. EVALUATION

In this section, we will provide a detailed analysis on encryption and generalization. The evaluation is made in terms of efficiency, complexity and security.

A. Efficiency and Complexity

The encryption algorithm discussed here is a simple and easy one, however this is implemented in order to make the procedure quicker with lesser time and space complexity as compared to Signature Scheme based on Discrete Logarithm and Probabilistic Encryption which are time consuming and may not be scalable. The encryption algorithm implemented in this paper is easier to compute, scalable and is also reliable as the keys are changed every time.

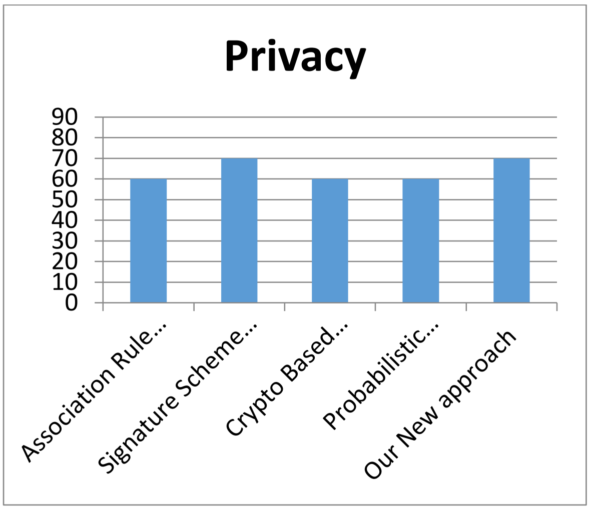
B. Privacy Analysis

Privacy of data is extremely high as in this paper even before encrypting the sensitive attributes the key itself is modified or rather encoded such that no eaves dropper or a brute-force attack can predict the values. Hence this ensures that the privacy of data is well taken care of and is doubled up as there is a two level encryption employed. Also some attributes may not require encryption however to protect the original values from being disclosed, the data is generalized.

6. EXPECTED RESULTS

Analysis was made on the current methodology with the required computations being applied to data sets taken from various fields. The methods proposed in this paper when compared to the methods presented previously prove to be efficient, fast, effective and scalable.

Encryption algorithm Results



7. CONCLUSION AND FUTURE WORK

We have proposed a novel approach of privacy preserving sensitive attributes through 2-level encryption technique and a generalization method which is secure, fast, less complex, robust and effective. Our future work will focus on encrypting numeric fields and extensive work will be conducted to assess the performance of the methods when the collaborated data is shared through some network.

REFERENCES

[1] S. Bhanumathi, Sakthivel, “A New Model for Privacy Preserving Multiparty Collaborative Data Mining” International conference on Circuits, Power and compting Technologies [ICCPCT-2013].

[2] Justin Zhanl, Stan Matwinl, Nathalie Japkowiczl, LiWu Chang,"Privacy-Preserving Collaborative Association Rule Mining," The Fourth International Conference on Electronic Business (ICEB2004) / Beijing, 2004.

[3] Elgamal,T, "A public key cryptosystem and a signature scheme based on discrete logarithms, Information theory," IEEE Transactions,Vol 31, No 4,JuI1985,pp.469-472.

[4] Justin Zhang,Stan Matwin, "A crypto-based approach to privacy- preserving collaborative data mining," Sixth IEEE International Conference on Data Mining-Workshops (ICDMW'06), Dec 2006, pp.546-550.

[5] S.Golwasser and S.Micali, "Probabilistic encryption," Journal of Computer and System Sciences, Vo1.28, pp.270-299, 1984.