A Review of Privacy Preservation Techniques

Gaurav Chauhan
GTU PG School, Gandhinagar

Abstract

In today`s heterogeneous connected environment privacy preservation is the biggest concern. Now a days each and every device, who can communicate, has some processing power and memory. The field of Privacy Preservation is explored by many people from cryptography to data mining to statistics. This area was mostly studied by data mining and cryptography people. The recent research in technology in the field of data mining has created renewed interest in the field of Privacy Preservation. The methods of data mining and retrieving became more sophisticated and one can retrieve data with more accuracy. In this situation it is necessary to have the basic idea about the techniques which are being used in Privacy Preservation field. In this paper I am trying to explore the techniques with various perspective and will try to give a combined idea about the work carried out by various authors from different communities.

1. Introduction

The problem of Privacy Preservation has emerged in various fields like data mining, mobile environment, sensor networks has shown very large growth. Due to new technologies with more power to store, process and communicate data, people started talking about misuse of data stored at various places. The researchers from different areas have seen this problem with their own perspective and tried to solve it with many techniques. Many research papers are available for Privacy Preservation Techniques. Now at this point it is necessary to prepare a survey of techniques which give basic idea about each technique and anyone who is willing to work can start from here. The techniques like Randomization, k-anonymity [1, 2, 3] etc. have been emerged as Privacy Preservation Techniques.

To start with we can identify following ways as Privacy Preservation Techniques:

- **Privacy Preserving using Transformation:** Here we use some transformation to hide sensitive data. Such transformation includes randomization [1], k-anonymity [4], and l-diversity [5]. In addition to this we have to keep in mind that the data being manipulated should be useful after processing.
- **Result Modification of Data Mining Applications for Privacy Preservation:** It is possible to retrieve the important information from data mining. So the basic idea is to modify the content of the result in such a manner that the privacy is not compromised. The technique called associative rule hiding is example of above idea.
- **Query Auditing:** These are similar technique to previously discussed technique. Here we supress or modify the result of query. Such techniques are discussed in [6, 7]

2. Privacy Preservation Using Transformation:

The straight forward solution for Privacy Preservation is transformation of sensitive data. The data is modified in such a way that sensitive data cannot be recovered. The price we pay is loss of effectiveness in data retrieval and mining as we have modified original data, this is obvious. Such techniques for Privacy Preservation are randomization [1], k-anonymity [4] and l-diversity [5].

2.1 Randomization

It is well known that if we add noise in the data it is hard to find actual data. This concept is employed in randomization technique. The sufficient large noise is added so that mining of sensitive data becomes impossible. By adding noise we mean that the attribute value of the record is masked [1, 2].

Mostly this method id used in public surveys where one can find evasive answer bias [8]. The method of randomization can be explained in such a way that on a data sets under consideration, the independent noise elements are added so the variance of noise is large enough that original data cannot be easily found.

After the randomization process the individual records are dissolved and we have one distribution who

has same behaviour of original data set. The real challenge is we have to modify existing data mining algorithm in such a manner that it can work on distribution rather than individual record [9].

One key advantage of randomization is that it is relatively simple and does not require the knowledge of other records. There for we don`t need to have secure server. Any system can randomize the data as all data is being treated equally. The weakness which may be exploited by attackers is that the data other than dense region is more susceptible [10].

One other method which is applied here in randomization is adding or dropping random items from data set. The results of this methods are shown here [11].

2.2 K-anonymity

While randomization was working on single individual elements, k – anonymity works on group. In anonymization important parameters related to identification of individual entity such as Unique Identity No (Aadhar No) is removed. But most of the times an individual entity can be identified by other identifiers like age, pincode and sex. These are known as pseudo identifiers. In k-anonymity [9] techniques we *generalize* or *supress* such pseudo identifiers.

In generalization we set the range of data sets and modify it. For example if we have presented the list of people from various cities like Jamnagar, Rajkot or Ahmedabad, we set the value of that column as Gujarat, the name of the state where these cities are located.

In suppression methods the sensitive information such as name of city from above discussed example is completely removed.

By using such methods we can achieve anonymization but the effectiveness of data is decreased.

In [12] we can see that the problem of k-anonymity is NP hard. The k-anonymity techniques requires that every tuple in the table is related to no fewer than k respondents in such a way that they are inseparable with other k columns.

The weakness here is if the attacker has the sample of data, the identification of original data becomes at risk. The more the sample the more risk of data. this knowledge will help to decide either use anonymization or not.

2.3 l-diversity

In k-anonymity we set value of sensitive data to some generalized value. Thus our data becomes anonymised, but the problem is such data is susceptible to attacks where background knowledge is available with the attacker. For example in *Homogeneity attack* attacker seeks the value which has exact same values in other words they are generalized. If attacker have some background knowledge the exact values of generalized values can be calculated. In other such case called *Background Knowledge Attack* an attacker tries to find relation between one or more identifiers to narrow down possible sensitive field [13].

So due to above flow in k-anonymity the l-diversity method is introduced. In l-diversity not only group of k is maintained but the diversity of information is also maintained. Here it should be noted that if we have n different kind of attributes we have to manage n diversity. So problem becomes challenging.

To further enhance this model the t-closeness model is introduced.in t-closeness method it is required that the distance between sensitive data should not be more than the threshold value t [14].

3. Result Modification of Data Mining Applications for Privacy Preservation:

Many time it is possible to retrieve information from the output of application. So there are techniques developed to modify the result for privacy preservation. One such technique is related to the concept of association rule mining is association rule hiding. The association rule hiding is achieved by distortion or blocking. In distortion we change the value of given transaction.in most of the cases we may work on binary values so we only need to change some bits. In other mode called blocking we leave some entries incomplete that incomplete entries prevents the association rules.

Here we note that it might possible that with sensitive rules we also loose non sensitive rules and there is possibility that some unknown or unwanted rules may be created. Such negative effects reduces the usage of data and gives inaccurate results.

The blocking method is NP-hard and the proof is given at [16]. The important factor for blocking is that it changes values to unknown rather than some dummy value (for example NULL).

The other technique is related to classification. Many time by using classifier attacker can have sensitive values. Here the main idea is to reduce the effectiveness of classifier. To reduce the effectiveness we modify data in such a way that classification becomes less effective. For example we can change sex or age group so particular group is not classified. We can find discussion of such downgrading process in [15, 16].

3. Query Auditing

In the cases where databases are not available for public access but there may be a public interface from which aggregate query is allowed. In such cases

some attacker can fire series of queries which can reveal sensitive information about data.

To prevent such attack we deny one or more queries from sequence of queries. We decide which query is to be denied and which to be not is decided upon the sensitivity of the data. The broad idea is available at [17, 18].

In Query auditing we deny queries which has overlapping results, such queries can work as potential threat to privacy of underlying data.Other than overlapping we use the denial based on size of the query. The query must satisfy the allowable size before execution.

In query auditing we can think two possible ways, one in which the sequence of queries is not known is called *online version* while in other case we know the sequence of queries known as offline version. A k-anonymity techniques guarantee certain level of privacy [19]. So we can use such techniques on the result of queries for preservation purpose. The other method for privacy preservation in query auditing is random sampling. The output of query is computed from random value of data sets so attacker cannot generate sets of query [20]. In another method we can add noise in the result of the query so the adversary never get sensitive data. We have to add small noise to achieve the privacy. In today`s era of extreme large database and Big Data it is always question that which of the above technique can be applied as well as relied upon.

Many times it is possible that data is not available at one place. The data is scattered at various places and the owner of data want to work on this data sets collectively. It is possible to use cryptography techniques for such applications. [8]

4. Limitations of Privacy: The Curse of Dimensionality

The above discussed methods works well when we have situation like the type of data we work is same kind, but unfortunately it never be. The curse of dimensionality says that if we have data with multiple dimensions the power of any technique is limited.

In [21] the authors shows the effects of increasing dimensionality with k-anonymity algorithm. Many time attacker has vast knowledge of data and thus methods such as l-diversity is proposed. To achieve high level of privacy many attributes have to be supressed that leads to loss of data.

The curse of dimensionality work with chaining effect. To achieve we supress records in database, but the information can be calculated by using identifiers in the database so by using methods such as l-diversity we also manage to diversify the attributes. But still there are chance of successful attacks. So we may repeat the above steps again. Finally which will lead us to very inefficient database. Thus if we go on with our algorithm we may find it in feasible or impossible at certain stage.

To stop at any given point is the trade off decision.

5. References

[1] Agrawal R., Srikant R. Privacy-Preserving Data Mining. ACM SIGMOD Conference, 2000.

[2] Agrawal D. Aggarwal C. C. On the Design and Quantification of Privacy Preserving Data Mining Algorithms. ACM PODS Conference, 2002.

[3] Samarati P., Sweeney L. Protecting Privacy when Disclosing Information: k-Anonymity and its Enforcement Through Generalization and Suppression. IEEE Symp. On Security and Privacy, 1998.

[4] Bayardo R. J., Agrawal R. Data Privacy through optimal k-anonymization. ICDE Conference, 2005.

[5] Machanavajjhala A., Gehrke J., Kifer D. l-diversity: Privacy beyond k-anonymity. IEEE ICDE Conference, 2006.

[6] Blum A., Dwork C., McSherry F., Nissim K.: Practical Privacy: The SuLQ Framework.ACM PODS Conference, 2005.

[7] Nabar S., Marthi B., Kenthapadi K., Mishra N., Motwani R.: Toward Robustness in Query Auditing.VLDB Conference, 2006.

[8] Warner S. L. Randomized Response: A survey technique for eliminating evasive answer bias. Journal of American Statistical Association, 60(309):63–69, March 1965.

[9] Silverman B. W.: Density Estimation for Statistics and Data Analysis. Chapman and Hall, 1986.

[10] Aggarwal C. C.: On Randomization, Public Information and the Curse of Dimensionality.ICDE Conference, 2007.

[11] Evfimievski A., Srikant R., Agrawal R., Gehrke J.: Privacy-Preserving Mining of Association Rules.ACM KDD Conference, 2002.

[12] Meyerson A., Williams R. On the complexity of optimal k-anonymity.ACM PODS Conference, 2004.

[13] Martin D., Kifer D., Machanavajjhala A., Gehrke J., Halpern J.: WorstCase Background Knowledge.ICDE Conference, 2007.

[14] Atallah, M., Elmagarmid, A., Ibrahim, M., Bertino, E., Verykios, V.:Disclosure limitation of sensitive rules,Workshop on Knowledge and Data Engineering Exchange, 1999.

[15] Chang L., Moskowitz I.: Parsimonious downgrading and decision trees applied to the inference problem.New Security Paradigms Workshop, 1998.

[16] Moskowitz I., Chang L.: A decision theoretic system for information downgrading.Joint Conference on Information Sciences, 2000

[17] Dobkin D., Jones A., Lipton R.: Secure Databases: Protection against User Influence.ACM Transactions on Databases Systems, 4(1), 1979.

[18] Kenthapadi K.,Mishra N., Nissim K.: Simulatable Auditing, ACM PODS Conference, 2005

[19] Samarati P.: Protecting Respondents' Identities in Microdata Release.IEEE Trans. Knowl. Data Eng. 13(6): 1010-1027 (2001)

[20] Denning D.: Secure Statistical Databases with Random Sample Queries.ACM TODS Journal, 5(3), 1980.

[21] Aggarwal C. C. Onk-anonymity and the curse of dimensionality.VLDB Conference, 200