Data Preserving Techniques for Collaborative Data Publishing

R.Indhumathi
PG Scholar, CSE,
MIET Engg.College ,
Tamil Nadu, India.

S.Mohana
Associate Professor,CSE,
MIET Engg.College,
Tamil Nadu, India.

Abstract

This paper addresses the issue of privacy preserving in data mining process. Specifically, while considering a scenario in which two parties owning confidential databases wish to run a data mining algorithm on the union of their databases, without revealing any unnecessary information. Here two types of attack are discussed and they are "insider attack" and "outsider attack" by colluding data providers who may use their own data records to infer the data records contributed by other data providers. This paper includes a formal protection model named k-anonymity, l-diversity and t-closeness. Also secure multiparty computation protocol and their relationship with privacy-preserving data mining are discussed.

Keywords: Privacy, Anonymization, SMC, Distributed database.

1. Introduction

Privacy preservation techniques are mainly used to reduce the leakage of formation about the particular individual while the data are shared and released to public. In order to do this, main thing done is not to disclose sensitive information. The data is modified and then published for further process. For this various anonymization techniques are followed and they are generalization, suppression, anatomization, permutation and perturbation [3]. By various anonymization the data is modified and it retains sufficient utility and that can be released to other parties safely. This whole process is called as privacy-preserving data publishing [1].

- The Classification of Attributes is done as Key attributes, quasi - identifiers (QI) and Sensitive attributes. Key attribute which is represented as Name, address, phone number which is uniquely identified and it always removed before release.

The Quasi-identifiers example, birth date, gender which can be used for linking anonymized dataset with other datasets. And the last one is Sensitive attributes example Medical records, salaries, etc and these attributes is what the researchers need, so they are always released directly.

Two types of information disclosure have been identified and they are [7]:

- Identity disclosure and attribute disclosure.

Identity disclosure occurs when an individual is linked to a particular record in the released table, such that attacker can easily identified from the release table. Attributed is closure occurs when new information about some individuals is revealed, i.e., the released data makes it possible to infer the characteristics of an individual more accurately than it would be possible before the data release. Identity disclosure often leads to attribute disclosure. Once there is identity disclosure, an individual is re-identified and the corresponding sensitive values are revealed. Attribute dis-closure can occur with or without identity disclosure.

Various algorithms are used in data mining for privacy preservation such as k-anonymity, l-diversity, t-closeness etc. And various protocols are used for preserving data from attackers.

2. The k-anonymity Method

K-anonymity requires each tuple in the published table to be indistinguishable from at least k-1 other

tuples. The idea in k-anonymity is to reduce the granularity of representation of the data in such a way that a given record cannot be distinguished from at least (k − 1) other records.

To prevent record linkage through QID, Samarati and Sweeney [1998a, 1998b] proposed the notion of k-anonymity, if one record in the table has some value qid, at least k−1 other records also have the value qid. In other words, the minimum group size on QID is at least k. A table satisfying this requirement is called k-anonymous [2]. In a k-anonymous table, each record is indistinguishable from at least k− 1 other records ith respect to QID. Consequently, the probability of linking a victim to a specific record through QID is at most 1/k.

Table1. 3-anonymous Inpatient Microdata

	Zip Code	Age	Nationality	Condition
1	1400*	<30	*	Heart Disease
2	1400*	<30	*	Heart Disease
3	1400*	<30	*	Viral Infection
4	1400*	<30	*	Viral Infection
5	1533*	≥40	*	Heart Disease
6	1533*	≥40	*	Cancer
7	1533*	≥40	*	Viral Infection
8	1533*	≥40	*	Viral Infection
9	1400*	3*	*	Cancer
10	1400*	3*	*	Cancer
11	1400*	3*	*	Cancer
12	1400*	3*	*	Cancer

In the given Table 1, where K=4 and QI={Zip code, Age, Nationality}, it provides example of K-anonymity. Therefore, for each of the tuples contained in the table T, the values of the tuple that comprise the quasi-identifier appear at least twice in T.

2.1 Attacks on k-Anonymity

Here two types of attacks are addressed and they are homogeneity attack and the background knowledge attack.

Homogeneity Attack: In this attack, all the values for a sensitive attribute within a group of k records are the same. Therefore, even though the data is k-anonymized, the value of the sensitive attribute for that group of k records can be predicted exactly.

Background Knowledge Attack: In this attack, the adversary can use an association between one or more quasi-identifier attributes with the sensitive attribute in order to narrow down possible values of the sensitive field further. An example one in which background knowledge of low incidence of heart attacks among Japanese could be used to narrow down information for the sensitive field of what disease a patient might have.
While k-anonymity protects against identity disclosure, it is insufficient to prevent attribute disclosure, so we go for next method, l-diversity.

3. l -diversity Method

While k-anonymity is effective in preventing identification of a record, it may not always be effective in preventing inference of the sensitive values of the attributes of that record. Therefore, the technique of l-diversity was proposed which not only maintains the minimum group size of k, but also focuses on maintaining the diversity of the sensitive attributes. Therefore, the l-diversity model for privacy is defined as follows:

Definition 3(l-diversity): Let a q* block be a set of tuples such that its non-sensitive values generalize to q*. A q* block is l-diverse if it contains l "well represented" values for the sensitive attribute S. A table is l-diverse, if every q* block in it is l-diverse.

The ℓ-diversity was proposed in [4] to prevent homogeneity attacks.

Table2. 3-Diverse Inpatient Microdata

	Zip Code	Age	Nationality	Condition
1	1405*	≤40	*	Heart Disease
4	1405*	≤40	*	Viral Infection
8	1405*	≤40	*	Cancer
10	1405*	≤40	*	Cancer
5	1533*	>40	*	Heart Disease
6	1533*	>40	*	Cancer
7	1533*	>40	*	Viral Infection
8	1533*	>40	*	Viral Infection
2	1406*	≤40	*	Heart Disease
3	1406*	≤40	*	Viral Infection
11	1406*	≤40	*	Cancer
12	1406*	≤40	*	Cancer

Now in Table 1, it's a 4-anonymous table, each tuple has the same values for the quasi-identifier as at least three other tuples in the table. Attacks on a k-anonymous dataset that allow an attacker to identify individual records. Defending against these attacks requires a stronger notion of privacy as ℓ-diversity which is given in Table 2, which prevents k-anonymity attack.

Let us define a q* block to be the set of tuples in

Table2 whose non sensitive attribute values generalize to q*. In spite of having background knowledge if there are ℓ "well represented" sensitive values in a q* block, then suspect needs ℓ −1 damaging pieces of background knowledge to eliminate ℓ−1 possible sensitive values and infer a positive disclosure. Thus, by setting the parameter ℓ, the data publisher can determine how much protection is provided against background knowledge, even if this background knowledge is unknown to the publisher.

The given table is said to be ℓ-diversified if every equivalence classes in the table contains at least ℓ well-represented sensitive attribute values. ℓ-diversity must guarantee that the SA value of a particular person cannot be identified unless the adversary has enough background knowledge to eliminate ℓ−1 SA values in the person's EC. Several measures were proposed to quantify the meaning of "well-represented" of ℓ-diversity. These include entropy ℓ-diversity [4], recursive (c,ℓ)-diversity [4] and simple ℓ-diversity [5,6].

3.1 Problems faced in l-diversity

- One problem with l-diversity is that it is limited in its assumption of adversarial knowledge. It is possible for an adversary to gain information about a sensitive attribute as long as they have information about the global distribution of this attribute.

- Another problem with privacy preserving methods in general is that they effectively assume all attributes to be categorical; the adversary either does or does not learn something sensitive. Of course, especially with numerical attributes, being close to the value is often good enough.

3.2 Attacks in l-diversity

Skewness Attack: When the overall distribution is skewed, satisfying l-diversity does not prevent attribute disclosure.

Similarity Attack: When the sensitive attribute values in an equivalence class are distinct but semantically similar, an adversary can learn important information. This leakage of sensitive information occurs because while l-diversity requirement ensures "diversity" of sensitive values in each group it does not take into account this mantical closeness of these values.

4. t-Closeness

Privacy is measured by the information gain of an observer. Before seeing the released table, the observer has some prior belief about the sensitive attribute value of an individual. After seeing the released table, the observer has a posterior belief. Information gain can be represented as the difference between the posterior belief and the prior belief. The novelty of our approach is that we separate the information gain into two parts: that about the whole population in the released data and that about specific individuals.

The t-closeness Principle:

An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. A table is said to have t-closeness if all equivalence classes have t-closeness.

The t parameter in t-closeness enables one to trade-off between utility and privacy. Now the problem is to measure the distance between two probabilistic distributions. There are a number of ways to define the distance between them. Given two distributions P = (p1, p2, ..., pm),Q = (q1, q2, ..., qm), two well-known distance measures are as follows. The variational distance is defined as:

$$D[\mathbf{P}, \mathbf{Q}] = \sum_{i=1}^{m} \frac{1}{2} |p_i - q_i|.$$

Here there are two probability distributions over the values and the distance between the two probability Distributions to be dependent upon the ground distances among these values. This requirement leads us to the Earth Movers distance (EMD).

The above fact entails that t-closeness with EMD satisfies the following two properties and they are Generalization Property and Subset property**.**

These two properties guarantee that the t-closeness using EMD measurement can be incorporated into the general Frame work of the Incognito algorithm [8].Suppose we have two sensitive attributes U and V. One can consider the two attributes separately, i.e., an equivalence class E has t-closeness if E has t-closeness with respect to both U and V.

5. Privacy for Collaborative Data Publishing

When data are gathered and combined from

different data providers, mainly two things are done, for anonymization process [9], [10] and they are:

- One approach is for each provider to anonymize the data independently and then they are aggregated (anonymize-and-aggregate, Fig. 1(a)), which results in potential loss of integrated data utility. In this model, the data providers hide the information for privacy and they provide to other data providers, so there will be loss in data utility.

- Another approach is collaborative data publishing[10],which anonymizes data from all providers as if they would come from one source (aggregate-and-anonymize, Fig. 1(b)), using either a trusted third-party (TTP) or Secure Multi-party Computation (SMC) protocols [10]. Here the data is aggregated first and then they are anonymized. The all data providers follow certain protocol or they trust the third party for privacy preserving.

Main goal is to publish an anonymized view of the integrated data, T*, which will be immune to attacks. Attacks are run by attackers, i.e., a single or a group of external or internal entities that wants to breach privacy of data using background knowledge, as well as anonymized data.

Collaborative data publishing is carried out successfully with help of trusted third party (TTP) or Secure Multi-Party Computation (SMC) protocols, that guarantees that the information or data about particular individual is not disclosed anywhere, the privacy is maintained with help of SMC and there will be better data utility. Here it is assumed that the data providers are semi honest. So certain protocols are set and the all data providers accept that protocol and they continue the process.

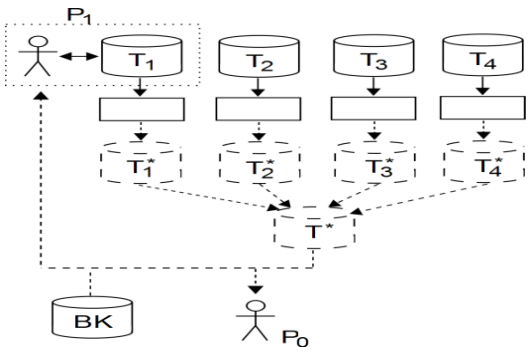

Figure 1(a). Anonymize-and-aggregate

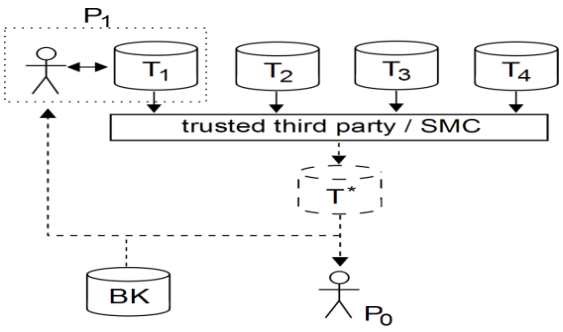

Figure1(b). Aggregate-and-anonymize

The m-adversary threats is explained with help of example, In Figure 2, Assume that bank P1, P2, P3, and P4 wish to collaboratively anonymize their respective customer databases T1, T2, T3, and T4. In general, multiple providers may collude with each other, hence having access to the union of their data, or a user may have access to multiple databases.

Here a new type of attack is identified by the data providers, which is "insider attack". And the data which is accessed by the hacker from outside is considered to be"outside attacker". Here the data is preserved from the inside attacker.

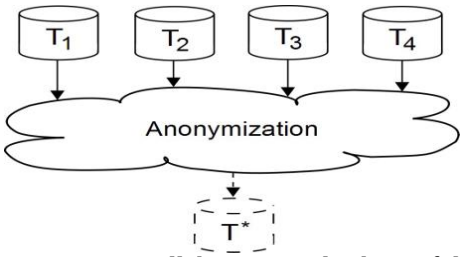

Figure 2.Collaborating 4 database of different providers

Figure3 describes how privacy is maintained for collaborative data. Here an m-adversary is defined as a coalition of m colluding data providers or data owners, and the attribute are further split and then the verification against the privacy constraint C is carried out to check whether the privacy is maintained or not.

After checking against C, if the attribute are able to further split ,then again the whole process is carried out from first. Finally the anonymized tabled is presented to m-adversaries. Where they cannot breach privacy of remaining records. Here pruning strategies are user to speedup the verification process.

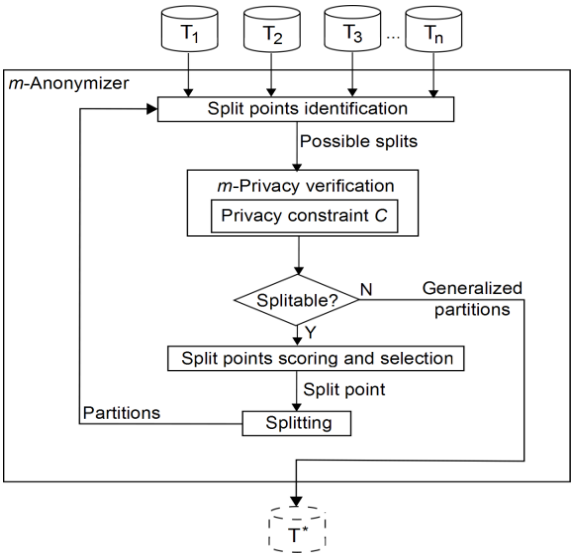

Figure 3.m-Anonymizer

Here three algorithms are used to define the whole process and they are:

- First, the notion of m-privacy is introduced, which guarantees that the anonymized data that is T* is checked against a constraint C, if it satisfies, that data providers can assume that privacy is preserved.
- Second, heuristic algorithms are used for exploiting the monotonicity of privacy constraints for efficiently checking m-privacy given a group of records.
- Third, data provider-aware anonymization algorithm is used for checking whether the data utility is high or not, whether data is efficiently used or not is checked here.

These are the process carried out for collaborating various database from different data providers.

7. Secure Multiparty Computation

When a distributed process is carried out, the secure multiparty computation is used. It is mainly used to control the malicious behaviour of data providers. SMC is mainly used to control the"insider attacker". The data providers are considered to be semi honest and they may try to verify the private record of other data provider, so to control this SMC is used. Thus, two important requirements on any secure computation protocol are privacy and correctness.

The SMC problems use two computation concepts: - Ideal model and Real model paradigm.

- In ideal model a Trusted Third Party (TTP) is used, which accepts inputs from all the parties, evaluates the common function and sends result of the computation to the parties. If the TTP is honest, then the parties can know the result only.
- In real model, there is no third party, instead all the parties agree on some protocol which allows them to evaluate the function while preserving privacy of individual inputs.

Secure computational Protocol for computation of sum of individual parties preserving privacy of their inputs. The protocol allows parties to break their data inputs into segments and distributing these segments among parties before computation.

8. Conclusion

Table3. Relationship between 3 Techniques

Techniques	Merits	Demerits
K-Anonymity	• Prevents linkage attack • Protects against identity disclosure.	•Can't prevent Attribute disclosure, Homogeneity and BK attack.
l-diversity	• Prevents Homogeneity and BK attack.	•Can't prevent Skewness and Similarity attack.
t-Closeness	• Overcomes demerits of k-Anonymity and l-Diversity. • Better in Utility and privacy.	• No computational procedure to reach t-Closeness.

From Table 3 it is concluded as, K-anonymity protects against identity disclosure, it is insufficient to prevent attribute disclosure. The ℓ-diversity was proposed to prevent homogeneity attacks. Main Intuition in l-diversity is the most frequent value does not appear too frequently. It also found that the data quality of k-anonymous tables without t-closeness is slightly better than k-anonymous tables with t- closeness. This is because t-closeness requirement provides extra protection to sensitive values and the cost is decreased data quality. All algorithms have been implemented in distributed settings with a TTP and as SMC protocols.

9. References

[1].Privacy-preserving data publishing an overview ,Author:Raymond Chi-Wing WongAda Wai-Chee Fu c2010.
[2].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.
[3].Karthikeyan.B,Manikandan. G,Vaithiyanathan. V," A FUZZY BASED APPROACH FOR PRIVACY PRESERVING CLUSTERING", Journal of Theoretical and Applied Information Technology,2011,Vol. 32 No.2.
[4].AshwinMachanavajjhala, Johannes Gehrke, DanielKifer, Muthuramakrishnan Venkitasubramaniam, ℓ-diversity: privacy beyond k-anonymity, IEEE International Conference on Data Engineering, 2006, p. 24.
[5].Xiaokui Xiao, Yufei Tao, Anatomy: simple and effective privacy preservation, International Conference on Very Large Data Bases, 2006, pp. 139–150.
[6].Gabriel Ghinita, PanagiotisKarras, PanosKalnis, Nikos Mamoulis, Fast data anonymization with low information loss, International Conference on Very Large Data Bases, 2007, pp. 758–769.
[7].G. T. Duncan and D. Lambert. Disclosure-limited data dissemination.J. Am. Stat. Assoc., pages 10–28, 1986.11. Copyright forms and reprint orders
[8].B. C. M. Fung, K.Wang, R. Chen, and P. S. Yu, Privacy-preserving data publishing: A survey of recent developments," ACM Comput. Surv., vol. 42, pp. 14:1‑14:53, June 2010.
[9].N. Mohammed, B. C. M. Fung, P. C. K. Hung, and C. Lee, "Centralized and distributed anonymization for high-dimensional healthcare data," ACM Trans. on Knowl. Discovery from Data, vol. 4,no. 4, pp. 18:1‑18:33, October 2010.
[10].Y. Lindell and B. Pinkas, "Secure multiparty computation for privacy-preserving data mining," The Journal of Privacy and Confidentiality, vol. 1, no. 1, pp. 59‑98, 2009.