Techniques of Privacy Preservation in Data Publishing

Nargis Nasir Sheikh¹, Prof. S. A. Bhavsar²

¹PG Student, MCOERC, Eklahare, Nashik.
²Assistant Professor, MCOERC, Ekalahare, Nashik.

Abstract—Privacy-preservation in data mining is used to safeguard susceptible information from unsanctioned disclosure. Due to the increasing ability to store personal data about users in privacy is an important issue involved data publishing. The different techniques such as bucketization, generalization have been proposed to perform privacy-preservation in data publishing. Recent work has shown that the technique generalization does not support for high-dimensional data. And also, bucketization cannot prevent membership disclosure and does not apply for data that do not have a clear separation between quasi-identifying attributes and sensitive attributes. The anonymization procedure, such as generalization along with bucketization, is actually intended regarding privacy conservation in microdata creating. In this paper, we presented different techniques of privacy preservation in data publishing. Here, we also study some application areas of privacy preservation in data publishing.

Keywords—Data Publishing, Privacy, Generalization, Bucketization, Data Anonymization.

I. INTRODUCTION

A. Privacy:

The word data privacy [1] refers to the relationship between collection and diffusion of data, ability, the public panorama of privacy, and the legitimate and opinionated issues adjoining them. Privacy apprehensions subsist wherever individually identifiable information is composed and stocked up – both in digital form or the other- wise. Indecent or unreal disclosure control can be the core cause for the privacy related issues. Data privacy concerns can arise in angry reply to information from a very wide range of resources, such as: Healthcare records, Residence and geographic records, Ethnicity, Criminal justice investigations and proceedings, Biological traits, financial institutions and transactions, Privacy Breach.

The difference of opinion in data privacy is to distribute data while shielding the personally identifiable type of information. The neighborhoods of data security and the information security work out and develop software, the hardware and also the human resources to deal with this issue. The privacy-preservation in data publishing model is shown in figure1.1

Figure 1.1 Privacy preservation in data publishing model

Data collection is the process of collecting the data form record owners (e.g., Alice and Bob). Data publishing is the process of proving collected data publically for data beneficiary. The privacy-preservation is the process of protecting publishing information from the attackers [2], [3].

B. Microdata Publishing:

In this paper, we consider microdata such as census data and medical data. Typically, microdata is stored in a table, and each record (row) corresponds to one individual. Each record has a number of attributes, which can be divided into the following three categories:

1. Identifier. Identifiers are attributes that clearly identify individuals. Examples include Social Security Number and Name.

2. Quasi-Identifier. Quasi-identifiers are attributes whose values when taken together can potentially identify an individual. Examples include Zip-code, Birthdate, and Gender. An adversary may already know the QI values of some individuals in the data. This knowledge can be either from personal contact or from other publicly available databases (e.g., a voter registration list) that include both explicit identifiers and quasi-identifiers.

3. Sensitive Attribute: Sensitive attributes are attributes whose values should not be associated with an individual by the adversary. Examples include Disease and Salary.

The field of privacy has seen rapid advances in recent years because of the increases in the ability to store data. In particular, recent advances in the data mining field have lead to increased concerns about privacy.
While the topic of privacy has been traditionally studied in the context of cryptography and information-hiding, recent emphasis on data mining has lead to renewed interest in the field. In section II of this paper, we will introduce the topic of privacy-preserving data mining and provide an overview of the different topics. Section III describes the application areas of privacy preservation in data publishing. The section IV concludes this paper.

II. LITERATURE SURVEY

This section elaborates different techniques used for privacy preservation in data publishing.

A. The Randomization Method:

The randomization method has been traditionally used in the context of distorting data by probability distribution for methods such as surveys which have an evasive answer bias because of privacy concerns [4, 5]. This technique has also been extended to the problem of privacy preserving data mining [6].

One key advantage of the randomization method is that it is relatively simple, and does not require knowledge of the distribution of other records in the data. This is not true of other methods such as k-anonymity which require the knowledge of other records in the data. Therefore, the randomization method can be implemented at data collection time, and does not require the use of a trusted server containing all the original records in order to perform the anonymization process.

1) Privacy Quantification:

The quantity used to measure privacy should indicate how closely the original value of an attribute can be estimated. The work in [6] uses a measure that defines privacy as follows: If the original value can be estimated with c% confidence to lie in the interval [a1, a2], then the interval width (a2 – a1) defines the amount of privacy at c% confidence level.

2) Adversarial Attacks on Randomization:

The higher the log-likelihood fit directly depends on the probability that the publicly public record (W) corresponds to record (X). If it is known that the public data set always includes X, then the maximum likelihood fit can provide a high degree of certainty in identifying the correct record.

3) Randomization Methods for Data Streams:

The randomization approach is particularly well suited to privacy-preserving data mining of streams, since the noise added to a given record is independent of the rest of the data.

However, streams provide a particularly vulnerable target for adversarial attacks with the use of PCA based techniques [7] because of the large volume of the data available for analysis.

4) Multiplicative Perturbations:

Multiplicative perturbations can also be used to good effect for privacy-preserving data mining. Many of these techniques derive their roots in the work of [8] which shows how to use multi-dimensional projections in order to reduce the dimensionality of the data. This technique preserves the inter record distances approximately, and therefore the transformed records can be used in conjunction with a variety of data mining applications. However, with some prior knowledge, two kinds of attacks are possible [9]:

Known Input-Output Attack: In this case, the attacker knows some linearly independent collection of records, and their corresponding perturbed version. In such cases, linear algebra techniques can be used to reverse-engineer the nature of the privacy preserving transformation.

Known Sample Attack: In this case, the attacker has a collection of independent data samples from the same distribution from which the original data was drawn. In such cases, principal component analysis techniques can be used in order to reconstruct the behavior of the original data.

5) Data Swapping:

A method data swapping, in which the values across different records are swapped in order to perform the privacy-preservation [10]. One advantage of this technique is that the lower order marginal totals of the data are completely preserved and are not perturbed at all. Therefore certain kinds of aggregate computations can be exactly performed without violating the privacy of the data.

B. Group Based Anonymization:

In [11], it has been shown that the use of publicly available records can lead to the privacy getting heavily compromised in high-dimensional cases. This is especially true of outlier records which can be easily distinguished from other records in their locality. Therefore, a broad approach too many privacy transformations are to construct groups of anonymous records which are transformed in a group-specific way.
1) The k-Anonymity Framework:

In k-anonymity techniques [12], we reduce the granularity of representation of these pseudo-identifiers with the use of techniques such as generalization and suppression. In the method of generalization, the attribute values are generalized to a range in order to reduce the granularity of representation. It is clear that such methods reduce the risk of identification with the use of public records, while reducing the accuracy of applications on the transformed data.

2) Personalized Privacy-Preservation:

Not all individuals or entities are equally concerned about their privacy. For example, a corporation may have very different constraints on the privacy of its records as compared to an individual. This leads to the natural problem that we may wish to treat the records in a given data set very differently for anonymization purposes. His approach has the advantage that it allows for direct protection of the sensitive values of individuals than a vanilla k-anonymity method which is susceptible to different kinds of attacks.

3) Utility Based Privacy Preservation:

The process of privacy-preservation leads to loss of information for data mining purposes. This loss of information can also be considered a loss of utility for data mining purposes. The generalizations performed on the marginal tables and the original tables in fact do not need to be the same. It has been shown that this broad approach can preserve considerable utility of the data set without violating privacy. Another direction for utility based privacy-preserving data mining is to anonymize the data in such a way that it remains useful for particular kinds of data mining or database applications.

4) Sequential Releases:

Privacy-preserving data mining poses unique problems for dynamic applications such as data streams because in such cases, the data is released sequentially. In other cases, different views of the table may be released sequentially. Once a data block is released, it is no longer possible to go back and increase the level of generalization.

5) The l-diversity Method:

The l-diversity is an attractive technique because of the simplicity of the definition and the numerous algorithms available to perform the anonymization.

Nevertheless the technique is susceptible to many kinds of attacks especially when background knowledge is available to the attacker. Some kinds of such attacks are as follows:

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.

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.

6) The t-closeness Model

The t-closeness model is a further enhancement on the concept of l-diversity. One characteristic of the l-diversity model is that it treats all values of a given attribute in a similar way irrespective of its distribution in the data.

7) Models for Text, Binary and String Data

These can be considered a case of text and market basket data, they are typically too high dimensional to work effectively with standard k-anonymization techniques. However, these kinds of data sets have the special property that they are extremely sparse. String Data is considered challenging because of the variations in the lengths of strings across different records.

C. Distributed Privacy-Preserving Data Mining

The key goal in most distributed methods for privacy-preserving data mining is to allow computation of useful aggregate statistics over the entire data set without compromising the privacy of the individual data sets within the different participants. Thus, the participants may wish to collaborate in obtaining aggregate results, but may not fully trust each other in terms of the distribution of their own data sets. For this purpose, the data sets may either be horizontally partitioned or be vertically partitioned. The problem of distributed privacy-preserving data mining overlaps closely with a field in cryptography for determining secure multi-party computations.
1) Distributed Algorithms over Horizontally Partitioned Data Sets:

In horizontally partitioned data sets, different sites contain different sets of records with the same (or highly overlapping) set of attributes which are used for mining purposes. A related problem is that of information retrieval and document indexing in a network of content providers. This problem arises in the context of multiple providers which may need to cooperate with one another in sharing their content, but may essentially be business competitors.

2) Distributed Algorithms over Vertically Partitioned Data:

For the vertically partitioned case, many primitive operations such as computing the scalar product or the secure set size intersection can be useful in computing the results of data mining algorithms. For the vertically partitioned case, many primitive operations such as computing the scalar product or the secure set size intersection can be useful in computing the results of data mining algorithms.

The approach of vertically partitioned mining has been extended to a variety of data mining applications such as decision trees [13], SVM Classification [14], Naive Bayes Classifier [15], and k-means clustering [16].

3) Distributed Algorithms for k-Anonymity:

The issue of k-anonymity is also important in the context of hiding identification in the context of distributed location based services. In this case, k-anonymity of the user-identity is maintained even when the location information is released. Such location information is often released when a user may send a message at any point from a given location. In many cases, it is important to maintain k-anonymity across different distributed parties.

A similar issue arises in the context of communication protocols in which the anonymity of senders (or receivers) may need to be protected.

III. Application Areas

The problem of privacy-preserving data mining has numerous applications in homeland security, medical database mining, and customer transaction analysis. Some of these applications such as those involving bio-terrorism and medical database mining may intersect in scope. In this section, we will discuss a number of different applications of privacy-preserving data mining methods.

Applications of Privacy-Preserving Data Mining:

A. Medical Databases:

The scrub system [17] was designed for de-identification of clinical notes and letters which typically occurs in the form of textual data. Clinical notes and letters are typically in the form of text which contains references to patients, family members, addresses, phone numbers or providers. Traditional techniques simply use a global search and replace procedure in order to provide privacy. The Scrub system uses numerous detection algorithms which compete in parallel to determine when a block of text corresponds to a name, address or a phone number. The Scrub System uses local knowledge sources which compete with one another based on the certainty of their findings.

The Datafly System [18] was one of the earliest practical applications of privacy-preserving transformations. This system was designed to prevent identification of the subjects of medical records which may be stored in multidimensional format. The multi-dimensional information may include directly identifying information such as the social security number, or indirectly identifying information such as age, sex or zip-code.

The generalizations in the datafly system are typically done independently at the individual attribute level, since the bins are defined independently for different attributes. The datafly system is one of the earliest systems for anonymization, and is quite simple in its approach to anonymization. A lot of work in the anonymity field has been done since the creation of the datafly system, and there is considerable scope for enhancement of the datafly system with the use of these models.

B. Bioterrorism Applications

In typical bioterrorism applications, we would like to analyze medical data for privacy-preserving data mining purposes. Often a biological agent such as anthrax produces symptoms which are similar to other common respiratory diseases such as the cough, cold and the flu. In the absence of prior knowledge of such an attack, health care providers may diagnose a patient affected by an anthrax attack of have symptoms from one of the more common respiratory diseases. The key is to quickly identify a true anthrax attack from a normal outbreak of a common respiratory disease. In many cases, an unusual number of such cases in a given locality may indicate a bio-terrorist attack. Therefore, in order to identify such attacks it is necessary to track incidences of these common diseases as well.
Therefore, the corresponding data would need to be reported to public health agencies. However, the common respiratory diseases are not reportable diseases by law.

C. Homeland Security Applications:

A number of applications for homeland security are inherently intrusive because of the very nature of surveillance. In [19], a broad overview is provided on how privacy-preserving techniques may be used in order to deploy these applications effectively without violating user privacy. Some examples of such applications are as follows:

1) Credential Validation Problem:

In this problem, we are trying to match the subject of the credential to the person presenting the credential. For example, the theft of social security numbers presents a serious threat to homeland security. In the credential validation approach, an attempt is made to exploit the semantics associated with the social security number to determine whether the person presenting the SSN credential truly owns it.

2) Identity Theft:

A related technology [20] is to use a more active approach to avoid identity theft. The identity angel system [20], crawls through cyberspace, and determines people who are at risk from identity theft. This information can be used to notify appropriate parties. We note that both the above approaches to prevention of identity theft are relatively non-invasive and therefore do not violate privacy.

3) Web Camera Surveillance:

One possible method for surveillance is with the use of publicly available webcams [18], which can be used to detect unusual activity. We note that this is a much more invasive approach than the previously discussed techniques because of person specific information being captured in the webcams.

4) Video-Surveillance

In the context of sharing video-surveillance data, a major threat is the use of facial recognition software, which can match the facial images in videos to the facial images in a driver license database. While a straightforward solution is to completely black out each face, the result is of limited new, since all facial information has been wiped out.

A more balanced approach [21] is to use selective downgrading of the facial information, so that it scientifically limits the ability of facial recognition software to reliably identify faces, while maintaining facial details in images. Thus, this approach has the flavor of a $k$-anonymity approach, except that it creates new synthesized data for the application at hand.

5) The Watch List Problem

The motivation behind this problem is that the government typically has a list of known terrorists or suspected entities which it wishes to track from the population. The aim is to view transactional data such as store purchases, hospital admissions, airplane manifests, hotel registrations or school attendance records in order to identify or track these entities. This is a difficult problem because the transactional data is private, and the privacy of subjects who do not appear in the watch list need to be protected.

IV. Conclusion

In this paper, we presented a survey of the different techniques of privacy-preservation in data publishing. We discussed a variety of data modification techniques such as randomization and $k$-anonymity based techniques. Also, methods for distributed privacy-preserving mining, and the methods for handling horizontally and vertically partitioned data. Finally, we talked about a number of diverse application domains for which privacy-preserving data mining methods are useful.

REFERENCES


