State-of-art in Statistical Anonymization Techniques for Privacy Preserving Data Mining

ALPA K. SHAH
Department of Master of Computer Applications,
Sarvajanik College of Engineering and Technology,
Surat, India-395001
alpa.shah@scet.ac.in

Abstract-**With the increased and vast use of online data, security in data mining has now become very important. Anonymity techniques have proved very useful in distributed computation. More techniques are still under research and improvements for achieving higher level of security in sensitive data. In this paper, we provide a review of the statistical Anonymization methods that can be applied for privacy preserving data mining. A brief evaluation is then made to point out the pros and cons of the same.**

Keywords: ***privacy-preserving, anonymization, perturbation, micro-aggregation, synthetic micro data***

I. Introduction

Today, with the increased storage of personal data of users on Internet, problem of privacy preserving data mining has become very crucial. Organizations and different agencies very often need to publish sensitive micro data, like medical data or census data or student information data for research purpose or for developers to develop applications. Decision makers and trend analyzer also need such type of micro data. The term micro data can be referred to as data about an individual, person, household, business or other entity. Micro data may be data collected by surveys, censuses or obtained from administrative records. This must be done in such a way that the confidentiality of the information provided by respondents is preserved. With these recent changes in trends, sensitive data is now easily available for malicious use. The main concern is that sensitive information should not be disclosed or misinterpreted. Mainly, these disclosures can be classified in two broad categories- Identity disclosure and attribute disclosure. Identity disclosure is when an individual entity can be distinguished from the published data. Inferring crucial information about the individual from the published data is attribute disclosure. At such stages, it is necessary to maintain confidentiality of the published data. Confidentiality can be termed as limiting data access and disclosure to authorized users and preventing access by or disclosure to unauthorized ones.

For example, consider the hospital data pertaining patients information The data contains attribute values which can uniquely identify an individual (pincode, nationality, age) or/and (name) and sensitive information corresponding to individuals (medical condition, salary, location). The organization wants to wants to publish in such a way that information remains practically useful and also identity of an individual is not compromised. In the example illustrated, an adversary may leak the information like Hindu Communal List from the published data. Anonymization techniques can be used here by either anoymizing fields like Name or remove direct variable like Religion. With the approaches of Anonymization Techniques various classified fields based on their importance can be protected.

#	Non-Sensitive Data				Sensitive Data	
No	Pincode	Age	Religion	Salary	Name	Condition
1	390005	58	Hindu	19000	Kumar	Heart Disease
2	390015	52	Hindu	12390	Raj	Heart Disease
3	390004	24	Christian	NA	Bob	Viral Infection
4	390018	78	Muslim	12890	Akhtar	Cancer
5	390075	26	Hindu	NA	Ravin	Thelsemia

Table -1 Hospital Records

#	Non-Sensitive Data			Sensitive Data
No	Pincode	Age	Religion	Condition
1	390005	58	Hindu	Heart Disease
2	390015	52	Hindu	Heart Disease
3	390004	24	Christian	Viral Infection

Table -2 Published Data

	Non-Sensitive Data			Sensitive Data
No	Pincode	Age	Religion	Condition
1	390005	58	Hindu	Heart Disease
2	390015	52	Hindu	Heart Disease

Table-3 Hindu Communal List

II. Statistical Anonymization Techniques

Anonymization is process of removing or modifying the identifying variables contained in the micro data dataset. Typically an identifying variable is one that describes a characteristic of a person that is observable, that is registered (identification numbers, etc.), or generally, that can be known to other persons. *Direct identifiers,* which are variables such as names, addresses, or identity card numbers. They permit direct identification of a respondent but are not needed for statistical or research purposes, and should thus be removed from the published dataset. *Indirect identifiers,* which are characteristics that may be shared by several respondents, and whose combination could lead to the re-identification of one of them. For example, the combination of variables such as district of residence, age, sex, and profession would be identifying if only one individual of that particular sex, age and profession lived in that particular district. Such variables are needed for statistical purposes, and should thus not be removed from the published data files. Anonymizing the data will consist in determining which variables are potential identifiers (this relies on one's personal judgment), and in modifying the level of precision of these variables to reduce the risk of re-identification to an acceptable level. The challenge is to maximize the security while minimizing the resulting information loss.

A. Methods based on data reduction

Data reduction methods are used to increase the number of individuals in the sample that share same kind of characteristics. Here, the techniques are devised to minimize the presence of unique identifiable records of individuals. Various techniques are employed under this which are briefly described below.

Removing variables

In this method direct identifiers from the data file are removed. It is desirable remove a variable only when it is highly identifying and no other protection methods can be applied. A variable can also be removed when it is too sensitive for public use or irrelevant for analytical purpose. For example, information on religion, diseases, etc. might not be released in a public use file while they might be released in a licensed file.

No	Pincode	Age	Name
1	390005	58	Kumar
2	390015	52	Raj
3	390004	24	Bob

Table 4: After Removing Disease Variable

Removing records

This is direct measure to remove the identifiable record from the data file. Its use is recommended only when the record is identifiable in spite of other protection techniques being employed. For example in census data, the record of a person suffering from a very rare disease (say 1 in million) may be removed rather than removing the variable for disease itself from the data file. Since removal of records largely impacts the statistical properties of the released data, removing records has to be avoided as much as possible.

Global recoding

The global recoding method consists in aggregating the values observed in a variable into pre-defined classes (for example, recoding the age into five-year age groups, or the number of employees in three-size classes: small, medium and large). The method applies to numerical variables, continuous or discrete. It affects all records in the data file. The method can also be applied to key variables (such as geographic codes) to reduce their identifying effect.

Top and bottom coding

This is a special case of global recoding which is applied to numerical or ordinal categorical variables. The variable like 'Salary' is common example. The highest values of this variable are usually very rare and therefore identifiable. Top coding at certain thresholds introduces new categories such as "monthly salary higher than Rs.150000", leaving unchanged the other observed values. The same reasoning applied to the smaller observed values defines bottom coding. When dealing with ordinal categorical variables, a top (or bottom) category is defined by aggregating the "highest" (or "smallest") categories.

Local suppression

Local suppression consists in replacing the observed value of one or more variables in a certain record with a missing value. The result is an increase in the frequency count of records containing the same (modified) combination. A criterion is therefore necessary to decide which variable in the risky combinations has to be locally suppressed. For example the published data from Table1 can be formed as:

No	Pincode	Religion	Salary	Name
1	390005	Hindu	19000	Kumar
2	390015	Hindu	12390	Raj
3	390004	Christian	12890	Bob
4	390018	Muslim	12890	Akhtar
5	390075	Hindu	12890	Ravin

Table: 5 After suppression of Salary variable.

B. Methods based on data perturbation

The methods employed here can be useful in two ways. First, if the data are modified, re-identification by means of record linkage or matching algorithms is harder and uncertain. Secondly, even when an intruder is able to re-identify a unit, he/she cannot be confident that the disclosed data are consistent with the original data.

Micro-aggregation replaces an observed value with the average computed on a small group of units (small aggregate or micro-aggregate), including the investigated one. The units belonging to the same group will be represented in the released file by the same value. When micro-aggregation is independently applied to a set of variables, the method is called individual ranking. When all the variables are averaged at the same time for each group, the method is called multivariate micro-aggregation.The easiest way to group records before aggregating them is to sort the units according to their similarity and the values resulting from this criterion, and to aggregate consecutive units into fixed size groups.

For example, consider the salary field of the micro data file. Here, salary in a particular range say Rs.5000 to Rs.10000 is fixed as Rs.7755. Now in all the records which have salary between Rs.5000 to Rs.10000, the value will be changed to Rs.7755. Such type of aggregation is termed as uni-variant micro aggregation. Here the sorting criterion is the variable itself.

Another type of technique more widely used is multivariate micro aggregation. It is based on combination of more variables. In previous case, a combination of Age variable can be used to compute values for micro-aggregations.

Data swapping

Data swapping was initially proposed as a perturbation technique for categorical micro data, and aimed at protecting tabulation stemming from the perturbed micro data file. Data swapping consists in altering a proportion of the records in a file by swapping values of a subset of variables between selected pairs of records (*swap pairs*). The level of data protection depends on the perturbation level induced in the data. A criterion needs to be applied to determine which variables and which records (the *swapping rate*) have to be swapped. For categorical data, swapping is frequently applied to records that are sample unique or sample rare, as these records usually present higher risks of re-identification. 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

Post-randomization (PRAM)

As a statistical disclosure control technique, post-randomization (PRAM) induces uncertainty in the values of some variables by exchanging them according to a probabilistic mechanism. PRAM can therefore be considered as a randomized version of data swapping. As with data swapping, data protection is achieved because an intruder cannot be confident whether a certain released value is true, and therefore matching the record with external identifiers can "easily" lead to mismatch or attribute misclassification. The method has been introduced for categorical variable but it can be generalized to numerical variables as well.

Adding noise

Adding noise consists in adding a random value ε, with zero mean and predefined variance σ^2, to all values in the variable to be protected. Generally, methods based on adding noise are not considered very effective in terms of data protection.

Resampling

Resampling is a protection method for numerical micro data that consists in drawing with replacement t samples of n values from the original data, sorting the sample and averaging the sampled values. Data protection level guaranteed by this procedure is generally considered quite low.

C. Generating synthetic micro data

Synthetic micro data are created using simulation algorithms to generate ‘real world’ data. Synthetic micro data are an alternative approach to data protection, and are produced by using data simulation algorithms. The rationale for this approach is that synthetic data do not pose problems with regard to statistical disclosure control because they do not contain real data but preserve certain statistical properties.

III. Discusssion and comparison on Anonymization Techniques

Data reduction techniques suffer from homogeneity attack. That is specially to mention that sensitive data lacks diversities in values. Also if adversary has additional background knowledge then he can infer sensitive data pertaining to individuals. During the application of anonymization techniques two important assumptions hold true. First, it may be very hard for the owner of a database to determine which attributes are or are not available in external tables. Second, a specific type of attack is assumed, but in real scenarios there is no reason why an attacker would not try other methods of attacks.

Perturbation techniques do apply independent treatment of different attributes, but the main disadvantage being reconstructing original data values back from the published data. Perturbation techniques also becomes vulnerable in **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. Also for **Known Sample Attack,** perturbation techniques are not found to be satisfactory. Here, 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.

Data swapping technique does not follow the general principle in randomization which allows the value of a record be perturbed independently of the other records. Therefore, this technique must be used with other techniques.

Generally, users are not keen to work with synthetic data as they cannot be confident of the results of their statistical analysis. Nevertheless, this approach can also help to producing “test microdata set.” In this case, synthetic data files would be released to allow users to test their statistical procedures to successively access “true” micro data in a data enclave.

IV. Conclusion

Person specific data in its original form often contains sensitive information. Malicious or adversaries are always there to disclose the sensitive information that is collected through various data collecting methods. Hence, there is always a need to preserve the private information of the individual or organizations. In this paper, various statistical methods of anonymization have been presented and then various challenges that they can pose against privacy preserving data mining are discussed. The further developments of these techniques would lead to an added advantage to solve complexities to preserve the privacy.

References

[1] A.Shamir. How to share a secret. Communications of the ACM, 22(11):612–613, November 1979.

[2] Chris Clifton and Donald Marks, Security and privacy implications of data mining, in Proceedings of the ACM SIGMOD Workshop o n Research Issues on Data Mining and Knowledge Discovery.

[3] Daniel E. O’Leary, Knowledge Discovery as a Threat to Database Security. In proceedings of the 1st International Conference on Knowledge Discovery and Databases(1991), 107-516.

[4] Turban and J.E. Aronaon. *Decision support Systems and Intelligent Systems*, Prentice-Hall, New Jersey, USA, 2001

[5] Grup Crises, Rovira i Virgili University of Tarragona, Dept. of Computer Engineering and Mathematics. Synthetic Microdata Generation for Database Privacy Protection. Research Report CRIREP-04-009, Sep. 2004

[6] Josep Domingo-Ferrer and Vicenç Torra, Ordinal, continuous and heterogeneous k-anonymity through microaggregation, *Data Mining and Knowledge Discovery*, vol. 11, Number 2, pp 195-212, September 2005. ISSN: 1384-5810

[7] L. Sweeney. Achieving *k*-anonymity privacy protection using generalization and suppression. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10 (5), 2002; 571-588.

[8] Luisa Franconi and Julian Stander, A model-based method for disclosure limitation of business microdata. Servizio della Metodologia di Base per la Produzione Statistica, Istituto Nazionale di Statistica, Rome, Italy. January 2001.

[9] Matthias Schmid1 and Hans Schneeweiss, 2005, The Effect of Microaggregation Procedures on the Estimation of Linear Models: A Simulation Study

[10] P. P. de Wolf, J.M.Gouweleeuw, P. Kooiman, L. Willenborg, Reflections on PRAM. Statistical data protection, proceedings of the conference, Lisbon, 1998.

[11] R. Brand, Microdata Protection through Noise Addition, Inference Control in Statistical Databases. From Theory to Practice . Lecture Notes in Computer Science, vol. 2316, Springer, 2002.

[12] R. J. Bayardo and R. Agrawal. Data Privacy through Optimal k-Anonymization. In *Proc. of ICDE-2005*, 2005

[13] Stephen Lee Hansen and Sumitra Mukherjee. A Polynomial Algorithm for Optimal Microaggregation

[14] Corsini, L. Franconi, D. Pagliuca, G. Seri, An application of microaggregation methods to Italian business surveys, Statistical data protection, proceedings of the conference, Lisbon, 1998.