An Inventive Approach for Data Privacy by Slicing

Amruta Sankhe[1] Neha Kunte[2] Sinu Mathew[3]
[1,2,3]Assistant Professor
[1,2,3]Department of Computer Engineering
[1,2,3]Atharva College of Engineering, University of Mumbai

Abstract— Data anonymization is the method of destroying tracks, or the electronic trail, on the information that would direct an eavesdropper to its origins. An electronic trail is the data that is left behind when someone transfer data over a network. This technique will not take away the original field design of the data being anonymised, so the information will still seem realistic in test data environments. Data Privacy has become the main focus of research in the world of data mining. There are number of techniques like, Bucketization and Generalisation which normally used in field of data privacy. But there are some disadvantages of using these techniques. In this paper we introduce technique called as Slicing which provides better privacy than above techniques and its able to handle high dimensional data.

Key words: Slicing, Data Privacy

I. INTRODUCTION

Privacy is becoming an increasingly significant issue in most data mining applications. This has provoke the development of a lot of privacy-preserving data mining techniques. A large portion of them use ran-domised data distortion techniques to mask the data for protect the privacy of sensitive data. [1]

The well-known privacy-preserved data mining mod-ifies obtainable data mining techniques to random-ised data. This approach include moderately the requirement of a besieged data mining task into the procedure of masking data so that required structure is maintained in the masked data [2]. The idea is simple but novel: it looks into the data generalisation concept from data mining as a way to hide detailed information, rather than locate trends and patterns. Once the data is masked, standard data mining methods can be applied without modification. The work demonstrated another positive use of data mining technology: not only can it determine useful patterns, but also mask concealed information. [2]

In this work, a novel technique is presented to build a classification set having both unlabelled and a small amount of labelled instances [3]. The model is built by using the Flow Classification Algorithm (FCA). The FC algorithm is proficient to judge internally on set of marked data. Before classification, the associ-ate set of attributes in the each record set are grouped using bucketization method. The superiority of models updated from them is sufficient for utilisation of unlabelled records, or whether more set of labelled records are needed for classification is processed. [3]

II. LITERATURE SURVEY

There have been many studies in the field on Split-ting of Data Anonymization techniques. This section presents a very brief review of the related and recent studies.

Generalisation is a data anonymization technique where real values are replaced with "less specific but semantically consistent values" [4]. Typically, numeric values are generalised into intervals and categorical values are generalised into a set of distinct values (e.g.,USA, Canada}) or a single value that represents such a set (e.g., North-America). Sensitive attributes which need not need to be displayed are converted to a semantic value, for ex. „*". [4]

Generalisation is one of the commonly anonymised approaches that replace quasi-identifier values with values that are less specific but semantically consistent. If at least two transactions in a group have distinct values in a certain column then all information about the item in current group is lost [4]. QID used in this process includes all possible items in the log. In order for generalisation to be effective, records in the same bucket must be close to each other so that generalising the records would not lose too much information [5]

The first, which we term bucketization [6], is to partition the tuples in T into buckets, and then to separate the sensitive attribute from the non-sensitive ones by arbitrarily permuting the sensitive attribute values within each bucket. The sanitised data then consists of the buckets with permuted susceptible values [6].For example, if the underlying table T, then the publisher might publish bucketization B. Of course, for added confidentiality, the publisher can completely mask the identifying attribute (Name) and may moderately mask some of the other non-sensitive attributes (Age, Sex, and Zip). [6]

Bucketizationon the other hand does not put off membership revelation and does not apply for data that do not have a clear separation between quasi-identifying attributes and sensitive attributes [7]. Bucketization does not prevent membership disclosure. Because bucketization publishes the QI values in their unique forms, an adversary can find out whether an individual has a record in the published data or not. Bucketization requires a obvious division between QIs and SAs. However, in many data sets, it is unclear which attributes are QIs and which are SAs [7]

Several anonymization techniques, like generalisation and bucketization, have been intended for privacy preserving micro data publishing. Current work has shown that generalisation loses significant quantity of information, particularly for

high-dimensional data. On the other hand, Bucketization does not prevent membership confession and does not apply for data that do not have a clear separation between quasi-identifying attributes and sensitive attributes [8]

One of the methods for privacy preserving data mining is that of anonymization, in which a record is released only if it is impossible to differentiate from k other entities in the data. In high dimensional space the data becomes thin, and the concept of spatial locality is no longer easy to define from an application point of view [9]. When the data contains a large number of attributes which may be considered quasi-identifiers, it becomes difficult to anonymise the data without an unacceptably high amount of information loss. This is because an exponential number of combinations of dimensions can be used to make precise inference attacks, even when individual attributes are partially specified within a range [9].

A brief yet systematic review of several anonymization techniques such as generalisation and bucketization, have been designed for solitude preserving micro data publishing. Recent work has shown that generalisation loses significant amount of information, especially for high dimensional data. On the other hand, bucketization does not prevent membership disclosure [10]. This paper focuses on effective method that can be used for providing better data utility and can handle high-dimensional data.[10]

III. EXISTING SYSTEM

Various micro data anonymization techniques have been proposed. The majority popular ones are generalisation for k-anonymity and bucketization for l-diversity. In both the way, attributes are partitioned into three categories [11]:

1) Some attributes are identifiers that can adversely point out an individual, such as Name or Social Security Number;
2) Some attributes are Quasi Identifiers (QI), which the opposer may already know (possibly from other publicly available databases) and which, when taken jointly, can potentially recognize an individual, e.g., Birthdate, Sex, and Zip code;
3) Some attributes are Sensitive Attributes (SAs), which are unfamiliar to the adversary and are considered sensitive, such as Disease and Salary.

In both generalisation and bucketization, one first removes identifiers from the data and then partitions tuples into buckets. The two methods diverge in the next step. Generalisation transforms the QI-values in each bucket into "less specific but semantically consistent" values so that tuples in the same bucket can-not be distinguished by their QI values. In bucketization, one detached the SAs from the QIs by randomly permuting the SA values in each bucket. [11]

IV. AIM

Our main aim is to create a web application which will provide a graphical user interface to let a user register and login into the web site. The user will be able to extract the data set which is to be sliced. Firstly, the user will perform Generalisation on the extracted data set, followed by bucketization. After performing bucketization, the user can perform multi set generalisation. Eventually, the user will be able to slice the extracted data set which will highly preserve the data. Note that, when the sliced data is presented to an unknown person, he/she will not be able to infer the original meaning of data hence thereby providing better security of the data.

V. OBJECTIVES

- To emulate and show the working of different anonymization techniques like generalisation, bucketization, etc.
- To implement an innovative approach called Slicing.
- To show how Slicing is better than the existing anonymization methods by considering a real time database

A. Advantages of Proposed System:

Slicing preserves better data utility than generalisation. It takes care of more attribute correlations with the SAs than bucketization. It can also hold high-dimensional data and data without a clear detachment of QIs and SAs. Slicing can be efficiently used for preventing attribute disclosure, based on the privacy requirement of l-diversity. We introduce a notion called l diverse slicing, which ensures that the adversary cannot learn the sensitive value of any individual with a probability greater than 1. Slicing covers up for the drawbacks of Generalisation and Bucketization thereby proving to be a better anonymization technique.

VI. METHODOLOGY

In this project consists of the following modules,

- Dataset Extraction
- Generalisation
- Bucketization
- Multi-Set Generalisation
- Slicing

1) Dataset Extraction - The dataset extraction module can be used to extract the dataset and it will be store-din the database for future use. Initially the dataset was selected, after that it will be split separate data and it can be stored in the table to the user database.
2) Generalisation - Generalisation module performs 2-anonymity process. In generalisation approach we use the identifiers data and Quasi Identifiers. Here the attribute age is Identifiers, and gender is Quasi Identifiers. The generalisation data can be retrieved from an original data. The dataset data"s are saved into two buckets.

3) Bucketization - Bucketization module can be per-forms 2-diversity process. In generalisation approach we use the Quasi Identifiers. Here the attribute work class is attributing. The bucketization data can be retrieved from an original data. The dataset data's are stored into two buckets.
4) Multi-Set Generalisation - Multi-set generalisation module performs 2-anonymity process. In multi-set generalisation approach we use the identifiers data and Quasi Identifiers. Here the attribute age is Identi-fiers, and gender, work class are Quasi Identifiers. The multi-set generalisation data can be recovered from an original data. The dataset's data are stored into two buckets.
5) Slicing - Slicing partitions the data set both vertically and horizontally. Slicing preserves better data utility than generalisation and can be used for membership disclosure protection.

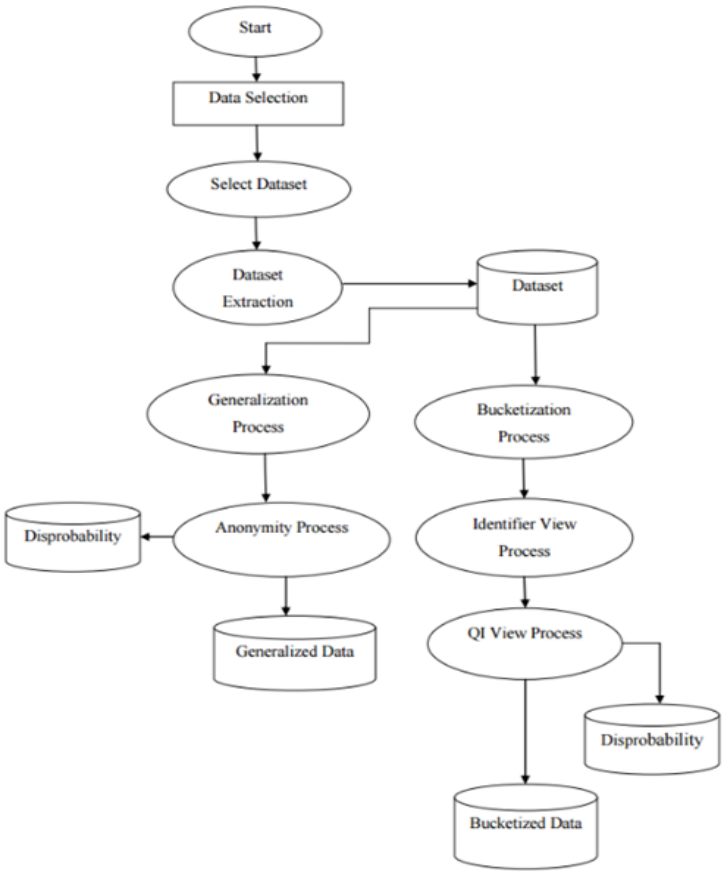

Fig. 1: Architectural Design

VII. CONCLUSION

Slicing is more proficient technique for data privacy than generalisation and Bucketization. It overcomes the boundaries of generalisation and Bucketization. As Bucketization, requires a partition of QI attributes and the sensitive attribute, slicing can be done with-out such a separation. The main ability of slicing is to handle high-dimensional data. Slicing is used to prevent attribute disclosure and membership disclo-sure.

REFERENCES

[1] Kargupta,H "On the privacy preserving properties of random data perturbation techniques", Data Mining, 2003. ICDM 2003. Third IEEE International Conference on 19-22 Nov. 2003 99 – 106 0-7695-1978-4 IEEE Dept. of Comput. Sci. & Electr. Eng., Univ. of Maryland Baltimore County, MD, USA
[2] Ke Wang; Yu, P.S.; Chakraborty, S., "Bottom-up generalization: a data mining solution to privacy protection," in Data Mining, 2004. ICDM '04. Fourth IEEE International Conference, vol., no., pp.249-256, 1-4 Nov. 2004 doi: 10.1109/ICDM.2004.10110
[3] G.Kesavaraj, S.Sukumaran "Bucketization based Flow Classification Algorithm for Data Stream Pri-vacy Mining", International Journal of Computer Applications (0975 – 8887) Volume 81 – No.12, November 2013
[4] Ji-Won Byun, AshishKamra, Elisa Bertino, and Ninghui Li, "Efficient k-anonymization Using Clus-tering Techniques" CERIAS and Computer Science, Purdue University, 2007
[5] Vasudha T, JanakiRamaiah B, "Sensitive Micro Data Disclosures Based on Tuple Grouping Meth-ods", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 12, December 2013
[6] Neha V. Mogre, Prof. GirishAgarwal, Prof. PragatiPatil "Privacy Preserving for Highdimensional Data using Anonymization Technique", Interna-tional Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 6, June 2013

[7] Mohd. Faquroddin, G KiranKumar,"A Better Approach for Privacy Preserving Data Publishing by Slicing", International Journal of Science and Re-search (IJSR), Volume 3 Issue 12, December 2014
[8] Vijay R. Sonawane, Kanchan S. Rahinj, "A New Data Anonymization Technique used For Member-ship Disclosure Protection", International Journal of Innovative Research in Science, Engineering and Technology, Vol. 2, Issue 4, April 2013
[9] C. Aggarwal, "On k-Anonymity and the Curse of Dimensionality,"Proc. Int‟l Conf. Very Large Data Bases (VLDB), pp. 901-909,2005.
[10] Amar P. Singh, Ms. DhanshriParihar, "A Review of Privacy Preserving Data Publishing Technique", International Journal of Emerging Research in Management and Technology, Volume 2 – Issue 6, June 2013.
[11] Dr. S. GovindaRao, D. Siva Prasad, M. Eswa-raRao "A New Approach Slicing for Micro Data Publishing", (IJCSIT) International Journal of Com-puter Science and Information Technologies