Perturbation of String Values

Krishna Priya .J   Geetha Mary. A
School of Computing Science and Engineering VIT University Vellore-632014, Tamilnadu, INDIA

Abstract- In the real time, large amount of repository data contain sensitive information that must be protected against unauthorised access. Hiding the sensitive data is one method, but may not be a best method since it explicitly denotes that sensitive values are hidden. Another method is removing the primary identifier and then anonymizing the data so that the individual records are protected. This method also explicitly denotes that information is hidden. We have gone for a crypt method, where sensitive data is transformed. Perturbation is the method used, which was basically used to convert numbers. Here we have proposed a frame work which uses perturbation methods to convert string data.

Keywords -Perturbation,Privacy preserving data mining, Randomization.

I. INTRODUCTION

In perturbation there are two approaches one is randomization and secure multiparty computation. Secure multiparty [9] computation is a problem which occurs with some value on a distributed network where each participant holds one of the inputs. In randomization random value is added to the sensitive attribute.

In perturbation there are two types, additive perturbation and multiplicative perturbation. Additive perturbation is done only horizontally but multiplicative perturbation is done both horizontally and vertically. It is mostly important in counterterrorism and homeland defence related applications and they may require for creating profiles and constructing social network models and detecting terrorist communications from privacy data.

II. RELATED WORK

Data mining will extract knowledge from various fields like marketing, weather forecasting, medical diagnosis and national security. In companies data are shared through data mining models to get a good set of customers and their buying habits.

It has data collection protocol and it will identify the necessary part of private information. It has some requirements like it must be scalable; it must be robust and low computational cost. There [9] are two methods of data collection protocols, Value based method and dimension based method. In value based method there are two approaches perturbation based approach and aggregation based approach. In dimension based method there are two approaches blocking based approach and projection based approach. In value based approach, a data provider manipulates value of each attribute. Perturbation based approach adds noise directly to the original data values such as changing age to 40 to 60. In aggregation based approach data will generalize according to exact domain level, such as changing age as 23 to 21-25. Dimension based approach removes the part of private information from the original data by reducing the number of dimensions. Projection based protocol will project only the original data and will have only the minimum information to construct accurate data mining models.

Randomization is one of the useful technique in privacy preserving data mining. There are two methods principal component analysis and Bayes estimate [1]. Let us take an example of age and for that age data are to be collected and then analysed both in size and complexity. In randomization scheme, a random number is added to the value of sensitive attribute. Principal component analysis method provides a framework to control the degree of redundancy. In bayes estimate we can able to solve the variable and use that variable as final reconstructed data.

Randomization has been used for variety of privacy preserving techniques. It is used in many fields such as statistics, computer sciences and social sciences. It is much importance to homeland security, modern science and to our real world i. e to the society.

Randomized transactions are very long and memory consuming. A large amount of information is specific to some individual users. Depending upon the nature of information, users may not be able to willing to divulge the individual values of records. Data mining techniques are considered a challenging task to privacy preservation due to natural tendency. Users are not equally protective of all values in their records. In many cases, data mining algorithms can be developed which use the probability distributions rather than individual records.

Large [3] amount of information is often specific to individual users. The privacy metric is based on the concept of mutual information between the original and perturbed records. There are two real facts, users are not equally protective of all users in the records, data mining algorithms can be developed which use the probability distributions rather than individual records.

Randomization has been a primary tool [4] to hide sensitive private information during privacy preserving data mining. It is only focused on how to reconstruct the distribution of the original data from the perturbed data.

Expectation maximization has low privacy loss and high [5] fidelity estimation of the distribution. This approach is not suitable for scenarios where a single party collects data from many users who don’t talk to each other and where this single party performs data mining operations on this data. Achieving privacy preservation [6] while sharing data for clustering is a challenging problem. Suppose that a hospital shares some data for research purposes. The hospitals security administrator may suppress some identifiers from patient records to meet privacy requirements. Two organizations, an internet marketing company and an online retail company have data sets with different attributes for a common set of individuals. In a horizontal partition, different objects are described with the same schema in all partitions. In vertical partition the attributes of same objects are split across the partitions.

Privacy is becoming an important issue in many data mining applications, that deal with health care, security, financial, behavioural and other types of sensitive data. Health organizations are allowed to release [7] data as long as identifiers are removed. Perturbation preserves statistical relationships between attributes, while providing adequate protection for individual confidential data. KD tree based perturbation [8] method, recursively partitions a data set into smaller subsets such that records within each subset are homogenous after each partition. Framework for mining association rules from data that have been randomized.

III . MEASURING THE PRIVACY

Privacy can be measured from the range from [- ∞ + ∞] and it is in terms of percentages.

By using differential equations we can measure the privacy[4]

Let us take the integer as

P(A)=- it lies from - ∞ + ∞ with log function log base 2

ie log $\int f(x)\,dx$ f(A)dA.

2nd order differentiation

P(A)=-2f(A) integral which lies from 0 to ∞ *1/f(A) dA

=-2 integral which lies from 0 to ∞*dA

=-2 ∞

Private information can be done in dependency attributes. Attributes may be simple or complex .attributes in many data sets are not independent and some attributes might have strong correlations among themselves. Data mining computations among several parties are performed on the combined data without revealing each party's data to other parties and solved using multiparty secure computation the random alphabet is tied to the data so that repeated queries return the same perturbed value. One of the design goals of such privacy preserving data mining is to derive efficient algorithms which can have a small privacy loss and a small information loss. Suppose that a retailer wants to make goods to the customer they must check that which product sales is very high. According to that they must manufacture the goods to meet the customer requirements. The sender and receiver will be able to send and receive the data.

Suppose that an online business arranges its webpages according to an aggregate model of its website visitors. In case of server, it has a complete and precise database with the information from its clients and it has to make a version of this database public, for others to work with. In particular, a company should not be able to match up records in the publicly released database, with the corresponding records in the company's own data base of its customers.

Random works can be explained by probabilistic frameworks. In randomization we have two metrics namely privacy loss and information loss to capture the amount of data in an individual record to the data mining algorithm. The key element in preserving privacy and confidentiality of sensitive data is the ability to evaluate to the extent of all potential disclosure for released data

III. QUANTIFICATION OF PRIVACY

The quantity used to measure privacy [4] should indicate how closely the original value of an attribute can be estimated. It is based on differential entropy of a random variable.

h(A)=$\int_{\square}^{\square}$ f base(a)log base 2fn(a) da.

V.Proposed Work:

We have proposed a frame work which converts sensitive string data using perturbation techniques.

Database: Database is a collection of information and it may in fields, records and files.

Field: A field is a single piece of information

Record: A record is one complete set of fields

File:A file is a collection of records

Extract Sensitive Data:

The data will be extracted sensitively and by database data will be extracted.

String to Number Conversion:

Text data is converted into an number

Perturbation:

Converts the number into an another number using one of the perturbation method.

Number to text conversion:

Again by using the inverse of the conversion method, which converts number to text data.

Perturbed String data :

converted text is displayed for the unauthorized users

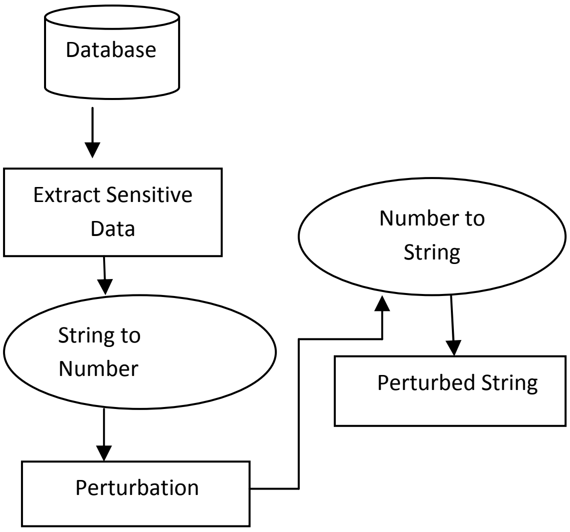

Fig: framework of string perturbation

VI.CONCLUSION:

Text is converted to numbers using the perturbation method and using the perturbation method we can convert the numbers to text and stored in the database.

ACKNOWLEDGEMENT:

I get this opportunity to express my sincere gratitude to my project coordinator and guide Asst .prof Geetha Mary.A ,School of Computing Science and Enginnering,Vellore Institute of Technology and further motivating and suggesting help and support for successful completion of this paper.

REFERENCES

[1] Zhengli Huang,Wenliang Du and Biao Chen,"Deriving Private information from randomized data" SIGMOD, Baltimore,Maryland,USA, 2005 June, Page14-16.

[2] Alexandre Evfimievski,Johannes Gehrke,RamaKrishnan SriKant,"Limiting Privacy Breaches in Privacy Preserving DataMining",PODS, San Diego,CA, June 2003, Page: 9-12.

[3] Dakshi Agrawal and Charu C.Aggarwal,"on the design and quantification of privacy preserving data mining algorithms SIGMOD, 2001, pages 439-453.

[4] Songtao Guo,Xintao Wu,"on the use of Spectral Filtering for Privacy Preserving Data Mining SAC'06 Dijon,France, April 2008, Pages 23-27.

[5] Chai Wah Wu,"Privacy Preserving data mining: a signal processing perspective and a simple data perturbation IBM Research Division", Thomas j.Watson Research Centre 2000

[6]Stanley R.M.Oliveira1,2 and Osmar, "achieving privacy preservation when sharing data for clustering ",Campinas,SP,Brasil, 2009.

[7] kun Liu,Hillol Kargupta, and Jessica Ryan,"Random projection based multiplicative data perturbation for privacy preserving distributed data mining", Maryland Baltimore country,Baltimore, IEEE Transactions on Knowledge and Data Engineering, January 2006.

[8]Xiao-Bai Li and Sumit Sarkar,"ATree based data perturbation for privacy preserving data mining", Proc Eigth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining,September 2007

[9] Nan Zheng and Wei Zhao," Privacy Preserving Data Mining Techniques," IEEE Computer Society, Cover feature ,April 2007.