A Review for privacy preserving using randomization for data mining

[1]Halak P. Patel, [2]Warish D. Patel

Department of computer science & engineering
Parul institute of Technology, Vadodara

ABSTRACT: **In many organizations large amount of data are collected. These data are sometimes used by the organizations for data mining tasks. However, the data collected may contain private or sensitive information which should be protected.Privacy protection is an important issue if we release data for the mining or sharing purpose. Our technique protects the sensitive data with less information loss which increase data usability and also prevent the sensitive data for various types of attack. Data can also be reconstructed using our proposed technique. A novel hybrid method to achieve k-support anonymity based on statistical observations on the datasets. Our comprehensive experiments on real as well as synthetic datasets show that our techniques are effective and provide moderate privacy.clustring based noise techniques that not only preserve the privacy but also ensure effective data mining.**

KEYWORDS: **Data mining, privacy preserving, k-anonymity, randomization.**

INTRODUCTION

Data mining and knowledge discovery in databases are two new research areas that investigate the automatic extraction of previously unknown patterns from large amounts of data. Data mining is concerned with the extraction of non-trivial, novel and potentially useful knowledge from large databases. Privacy preservation one of the most important issues in data mining. The privacy preserving mining methods modify the original data in some way, so that the privacy of the user data is preserved and at the same time the mining models can be econstructed from the modified data with reasonably accuracy. Various approaches have been proposed in the existing literature for privacy-preserving data mining which differ with respect to their assumptions of data collection model and user privacy requirements.

In this paper,focus on non-cryptographic techniques.privacy preserving used k-anonymity algoritham algorithm. When micro data is released for the research purpose, one needs to limit disclosure risk while maximize the utility of data. Sweeny introduced the k-anonymity technique to limit the disclosure risk . K - anonymity requirements says that, a data set is k anonymous (k 2: I) if each record in the data set is indistinguishable from at least (k-l) other records within the same data set. This k-anonymity requirement is generally achieved by using generalization and suppression . In generalization the attribute values are generalized in a particular interval. In suppression the attribute values are replaced or modified with some other values. Suppression contains information loss so it is generally avoided.

Key Attribute	Quasi identifier			Sensitive Attribute
ID	Sex	Age	Zip code	Disease
1	M	20	13000	Flu
2	M	24	13500	HIV+
3	F	26	16500	Fever
4	F	28	16400	Cancer

Table 1 :Diagnosis Data Set

Key Attribute	Quasi identifier			Sensitive Attribute
ID	Sex	Age	Zip code	Disease
1	M	[20-24]	13*00	Flu
2	M	[20-24]	13*00	HIV+
3	F	[26-28]	16*00	Fever
4	F	[26-28]	16*00	Cancer

Table 2:anonymity view of table 1

k-anonymous table contain three types of attributes. First is key attributes like name, SSN No, ID etc. which can be used toidentify the individuals uniquely. Second is quasi identifier attribute which are generally linked with publically available database to re-identify the individuals. This is called linking attack [8]. Third are sensitive attribute which needs to be protected. In table I we see the diagnosis data set. Table II shows the 2-anonymous view of table I.

Randomazation In random perturbation the privacy of data can be protected by perturbing the sensitive data with randomization algorithm before releasing it to the data miner [9]. The original data is distorted through adding the noise component to the data which is obtained through randomization. This method [1 0] deals with character type, classification type, boolean type and number type of discrete data. To facilitate the conversion of data sets, it is necessary to preprocess the original data set. This paper uses the method of average region to disperse the continuous data. Discrete formula is as follows: A (max) - A(min)/n = length. A is continuous attributes, n is number of discrete, length is the length of discrete intervals.

Main motive of this research work are:
Data Utility: Data utility is an important part because if data utility is minimum then it's also affects the accuracy of data mining tasks. Our goal is to eliminate the privacy leaks and increase the data utility.
Privacy: The core concept of this research work is privacy. Sensitive values are needed to be secured and for this we proposed a double layer security in the data set using randomization and k - anonymity.. information loss: Information loss should be minimized. This paper proposed a modified approach in k -anonymity technique which helps to keep the information loss minimum.

RELATED WORK

To the best of our knowledge, the work of Wong and Lim [12] is the first encryption solution to PPFIM. Here the data is encrypted using substitution cipher and random fake items are added to the transformed database.

The data owner using a one-to-n mapping encrypts the original transaction database to be sent to the third-party miner. The data miner performs the datamining task on the encrypted database and hands over the association rules to the data owner.

The data owner then extracts the original rules from the encrypted rules by decryption. The flaws in this methodology have been identified in [4] by proposing a frequency analysis based method for breaking the encryption scheme. The main flaw in the former work was that the fake items added in the encrypted database were independent of other items.

Gianotti et al [5] adopt a frequency-based attack model where the adversary knows the exact set of items along with the item support. Even in this work, the approach of k-support anonymity named differently by the authors as k-privacy is undertaken. The work considers adding fake transactions containing of real items for achieving k-privacy.

Providing privacy-preserving for Internet data is a longstanding goal of the computer research community. It is has received considerable attention with the development of data mining and network Technology.Cryptograph based privacy-preserving method can provide a better guarantee of the privacy when different institute want to cooperate in a common goal[4].

Noise addition using random perturbation technique has been explored in [2]. However since it uses random perturbation technique therefore even though privacy is obtained, data mining is affected since noise is added randomly without identifying the data characteristics.

T closeness model [1] uses the k-anonymity and 1 diversity approach but in addition ensures that that the distance between the distribution of the sensitive attribute within an anonymized group should not be different from the global distribution by more than a threshold t. Compared to the previous methods.this model provides better privacy but there is information gain for the attacker and data characteristics are also lost.

CONCLUSION

In this paper, they proposed a privacy preservation search scheme to enable accurate, efficient and secure search over encrypted private data. The proposed Hybrid approach employs randomization and K-anonymity method. By using Randomization technique attacker cannot identify a pattern of data. K-anonymity method has shortcoming of homogeneity and background attack. In the proposed method we combined K-anonymity with randomization. It makes difficult for the attacker to identify background and homogeneity attack. Apart from that it protects private data with better accuracy and gives no loss of information which increases data utility. Data can also be reconstructed by our proposed approach.

REFERENCES

[1]“An Efficient Approach for Privacy Preserving in Data Mining”Manish Shannal Atul Chaudhar/ Manish Mathuria3 Shalini Chaudhar/ Santosh Kumar5, IEEE , 2014

[2] “Privacy Preservation Algorithm in Data Mining for CRM Systems”Shashidhar Virupaksha, G Sahoo, Ananthasayanam Vasudevan,2014, IEEE

[3] “Dataless Data Mining: Association Rules-based Distributed Privacy-preserving Data Mining” Vikas G. Ashok, IEEE 2015

[4] “Privacy-Preserving Frequent Itemset Mining in Outsourced Transaction Databases” Iyer Chandrasekharan P.K. Baruah , 2015, IEEE

[5] **“**SYMMETRIC-KEY BASED PRIVACYPRESERVING SCHEME FOR MINING SUPPORT COUNTS”,Yu Li1 AND SHENG ZHONG, IEEE, 2013

[6] “K-Anonymity for Privacy Preserving Crime Data Publishing in Resource Constrained Environments”, Mark-John Burke, Anne V.D.M. Kayem, IEEE, 2014