A STUDY OF PRIVACY PRESERVATION IN DATA MINING

[1]G.Sowmya , [2]B.Vamshi , [3]G.Navya , [4]Dr. K.V.Naganjaneyulu
[1]STUDENT, Dept. of CSE, BIET, sowmyagonuguntla46@gmail.com
[2]STUDENT, Dept. of CSE, BIET, vamc96@gmail.com
[3]STUDENT, Dept. of CSE, BIET, navyagonuguntla1@gmail.com
[4]PROFESSOR, Dept. of CSE, BIET, kvn.dr@biet.ac.in

Abstract:

In order to extract the appealing unfamiliar patterns from large data sets, the most commonly used technique is Data Mining. The condition to preserve the privacy of data should be satisfied during the transmission of data to the third parties. From the continuous data records, extracting the knowledge structures is defined as Data Stream mining. Emerging data is an important problem in the Stream data mining. Privacy of the individual data, without losing the accuracy is dealt with Privacy preserving data mining abbreviated as PPDM. For the business decision making and evaluation, data is an important aid and on the other hand, lot of privacy concerns arise which prevents the owners of the data to share the particular information for the purpose of Data analysis. The Data Mining task called Clustering and Classification helps to achieve this certainty and Privacy measure. A competent and efficient approach which aims the delicate information's privacy and the data access with least information loss has been proposed. It considers the usage of Min-Max normalization and the addition of noise to the actual data which is a compound method in order to conserve the data privacy.

Keywords: *Data Stream, Data Mining, Classification and Clustering, Privacy preservation, Min-Max Normalization*

I. INTRODUCTION:

These days, the continuous data collection capability was facilitated in the Hardware Technology. Usage of the Debit and Credit cards which are the basic transactions of day to day life, leads to automated data storage. Large volume of data generally tends to mining and computational challenges.

All the fields like sales finance or marketing uses the concept of Data mining. Besides, the accelerated improvement in the communication technology and Internet led to Data streams. In the same way, many companies experience the loss of privacy of the data due to the frequent exposure to the data analysis.

Dynamic updating is required by the traditional Privacy Preserving data stream mining environment. For an instance, the traditional method's execution efficiency cannot be expanded to the user demand for the large amounts of incoming data any longer. Also the confined memory space and countless number of data streams have limited the accuracy of the mining result. Hence, in the recent years "Privacy Preservation in the data stream mining" has become an important issue in the Data Mining due to the reasons mentioned above.

In recent times, the emerging new type of data which is different from the static data is called Data stream. Following are the various characteristics of Data streams:

- Amount of data is infinite and continuous.
- It requires fast and real time response.
- Information distribution changes with respect to the time.

For the static Databases, traditional data mining algorithms are not designed. It is always necessary to rescan the database again and again if the data is changed. This leads to inability to prompt the user request and also computational delay. Various Privacy preserving techniques exists where a matrix based spectral filtering technique is proposed for the recovering of the original data from the irregular data. Based on the data correlations, the two data

reconstruction techniques are proposed. Principle component analysis abbreviated as PCA is used by one method and Bayes estimate technique is used by the other. This analysis shows that, the original data is reconstructed with ease and accuracy when there is a high correlation between the data attributes.

In order to preserve the sensitive information in the data, Perturbation technique is involved where the original data is perturbed i.e. by the addition of noise to the original data.

Generalizing across the population is the main idea of Data mining rather than revealing the information about individuals. But, evaluation of the individual data is the main problem with the working of Data mining

Need of privacy in data mining

Constant usage of various activities like Credit cards, debit cards, usage of emails and phones leads to large number of computerized trails. Ideally with the consent of data subjects, all the particular data must be collected and proper assurance of the privacy should be provided to the individuals by the collector.

II. PRIVACY PRESERVING IN DATA MINING

Due to the colossal benefits of Data mining, at the moment, there is a lot of demand for Privacy preserving data mining techniques due to the high public concerns. Extraction of the useful information along with the individual privacy is provided by the privacy preserving data mining technique.

Privacy preserving Data mining technique is enabled using various methods. For the sake of protecting the individual records from being re-identified, data set is modified before its release. When the data set is modified, an intruder cannot be certain about the re-identified correctness of data. This principle is used in one of the particular class of the various methods.

A favorable Privacy preserving technique must satisfy two main requirements called Privacy and High data quality. The perturbed data set's privacy and quality of data has to be evaluated.

III. LITERATURE SURVEY

The study of privacy preserving data mining techniques started extensively, covering the development approximately in two categories: Perturbation – Base technique and Secure – Multiparty computation base technique. The main idea of perturbation – based technique involves increase of noise to the raw data in order to perturb the original data distribution and preserve the content of the hidden raw data.

There are many types of methods for protecting the numerical data from disclosure. This consists of sampling, local suppression, random noise rounding and micro – aggregation.

There are different masking techniques that are very important to protect the sensitive data. Masking techniques are used to prevent the confidential information in the table. These techniques can be operated on different data types. Data types can be categorized as follows:

- Continuous Variables.
- Categorical Variables.

A. Continuous Variables:

This are also referred as cardinal, metric and scalable variables. The differences between the values are meaningful so that the arithmetic operations are performed.

B. Categorical Variables:

This are also referred as non – metric variables. The values are set of categories and standard arithmetic operations cannot be performed. There are two different types of categorical data they are as follows:

- Nominal Variables.
- Ordinal Variables.

Masking techniques are classified into two different categories:

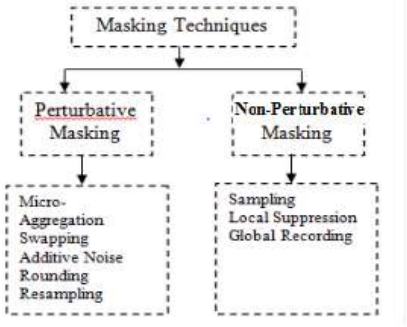
– Perturbative:

The original data are modified.

– Non-Perturbative:

The original data are not modified but some data are suppressed and some details are removed.

The figure below shows the classification of different masking techniques:



Geometric data transformation method (GDTM's) is one of the simple and typical example of data perturbation technique, which perturbs the numeric data with confidential attributes in cluster mining in order to preserve the privacy. Kumari et al proposed a privacy preserving lustering technique of fuzzy-sets, transforming confidential attributes into fuzzy items in order to preserve privacy.

Further some the largest issue encountered when implementing a perturbation technique is the inaccurate mining results from a perturbed data.

In view of this issue, the technique of random data perturbation introduced and this technique derives the original data distribution using a random noise for the data distribution and constructs a result similar to the original data.

Among the cluster mining algorithms, K-Means is one of the most popular and well-known methods mainly used due to its simple concept, easy implementation and comprehensible mining result.

C. Normalization Techniques for Privacy preserving in data mining:

In they have described the use of different normalization technique like Min-Max normalization Z-Score normalization and Decimal- Scaling methods with respect to privacy and accuracy, K–Means clustering algorithm is applied to the original data and the tailored data to verify the effectiveness and the correctness of the data . Here min – max normalization is used for preserving privacy during the mining process. The original data is sanitized using the min–max normalization approach before publishing.

The purpose of Normalization techniques is to map the data to a diverse scale. Various types of normalization techniques are available and they have compared the following normalization techniques - Min–Max normalization, Z – Score normalization and Decimal Scaling Normalization.

1) Min–Max Normalization:

Min–Max Normalization performs a linear alteration on the original data. The values are normalized within the given range. For mapping a v value an attribute A from range [min_A , max_A] to a range [new_minA , new_maxA], the computation to evaluate v' is given by

$$\frac{v - min_A}{max_A - min_A} (new\_max_A – new\_min_A) + new\_min_A$$

Where v' is the new value in the required range. The benefit of Min – Max normalization is that all the values are annealed within the certain range.

2) Z – Score Normalization:

Z – Score normalization is also called as Zero mean normalization. Here the data is normalized based on the mean and the standard deviation. Then the required formula to compute the result is:

$$d' = d - \frac{mean\ (P)}{std\ (P)}$$

Where Mean (P) = Sum of all attribute values in P.

Std(P) = Standard Deviation of all values of P.

3) Decimal Scaling Normalization:

Decimal scale normalization is based on the movement of the decimal values of attribute. The decimal point are moved depends on the maximum lute value of the attribute. The decimal scale normalization formula is:

$$d' = \frac{d}{10m}$$

Where m is smallest integer such that max (|d'|)< l.

IV. EVALUATING PRIVACY PRESERVING ALGORITHMS

Evaluation of Privacy preserving algorithms against certain parameters is another important aspect. The parameters are described as:

– Data utility: Loss in the functionality of data in providing the results is basically measured in

Data utility. It can be generated in the absence of PPDM algorithms.

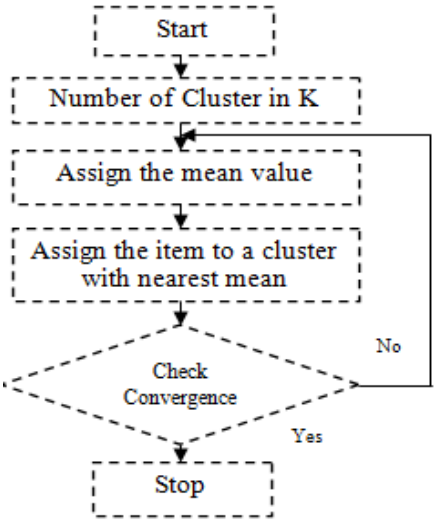
- Performance: The time taken to achieve the privacy criteria helps in measuring the performance of a mining algorithm.
- Uncertainty level: Prediction of the sensitive information that has been hidden is the measure of – uncertainty.
- Resistance: PPDM algorithm's tolerance against various data mining models and algorithms is measured in this context.

Information loss and quantification of privacy are the two important criteria. The measure which indicates how closely the original value of an attribute can be estimated is called the Quantification of privacy. The privacy is low if Quantification of privacy is estimated with higher confidence and vice-versa.

K-means clustering algorithm:

Grouping of similar object to appropriate clusters is defined as clustering which is an un-supervised learning method.

The flowchart shown below summarizes the K-mean algorithm steps:



V. CONCLUSION:

In a view to protect the sensitive data, privacy is the main approach. The sensitive information which doesn't want to be shared is worrying people the most. Existing techniques which are already present in the field of Privacy preserving Data mining are mainly focused in our survey. From the analysis we carried out, a single technique is not used for all the domains.

Depending upon the type of data, applications or domains, all methods are performed in a different way. However from our analysis, we concluded that rather than all the other methods, Random Data Perturbation methods and Cryptography perform better. For the encryption of delicate and sensitive data, cryptography is the best technique.

Additionally, high sensitivity of data is achieved with the help of Data perturbation which helps to preserve data. In the end, the level of privacy is made efficient by perturbation technique with normalization. Hence, than all the existing techniques, perturbation with normalization is the most important technique.

REFERENCES

[1] Shameemul Hague & Prince Shoeb Khan "Privacy Preserving in the Data Mining by Normalization". International journal of Computer Application (IJCA) Vol 96 No 6 Jun 2014, PP: 14-18.

[2] Gupta, and I. Rajput, "Preserving Privacy Using the Data Perturbation in Data Stream. (IJARCET), Vol. 2, No 5, May 2013.

[3] Kumari P, Raju K "Privacy preserving in cluster-ing using Fuzzy Sets". Proceedings of the 2006 International Conference on Data Mining, Las Vegas, Nevada, USA 2006, PP: 26 – 29

[4]. Yogita, D. Toshniwal "Clustering Techniques for Streaming Data-A Survey", 3rd IEEE International Advance Computing Conference (IACC), pp.951-956, 2012

[5] D.Patil,T.N.Rashmi & S.M Akhtar Authentication in Data Mining" International Conference on Advances in Computer and Electrical Engineering (ICACEE), pp.59-63, Nov 2012.