A Survey on Privacy Preservation in Data Mining

Alpesh Iyer[1] Ms Jasmine Jha[2]
[1]Student of M.E
[1,2]Department of Computer Science & Engineering
[1,2]L.J. Institute of Engineering & Technology

Abstract— Data mining is most usually used technique to extract the unfamiliar patterns from large data sets . Any transmission information to third parties i.e they should satisfy the condition to preserve the privacy of the data . Data Stream mining is the process of extracting knowledge structures from continuous data records[5]. The main problem in stream data mining is the evolving data . Privacy preserving data mining (PPDM) deals with the privacy of the individuals data and also without loosing the utility(accuracy) of the data[16] . On one hand data is the important asset for business decision making and analyzing on it . At the same time on the other hand the same data has many privacy concerns that might prevent the data owners to share those information for data analysis .This privacy & accuracy measure can be achieved by data mining task – Clustering & Classification . An efficient and effective approach has been proposed which aims at the privacy of the sensitive information and obtaining data with minimal information loss [14] . By using the Min-Max Normalization and by adding noise to the original data which is used as the composite method to preserve the privacy of the data[18] .

Key words: *Data Stream, Data Mining , Classification & Clustering , Privacy preservation , Data Perturbation , Min-Max Normalization*

I. INTRODUCTION

Nowadays, in hardware technology it has facilitated the ability to collect the data continuously. Simple transaction of everyday of life like by using debit cards, credit cards it leads to automated data storage. When the volume of the data is very large then generally it leads to number of computational and mining challenges [5].

Data mining is often used in marketing , sales and finance . Besides this , the rapid enhancement and the usage of internet and communication technology has led to data streams [18] . Similarly there are many companies which frequently expose their private data for data analysis purpose .which leads to the loss of privacy of the data .

Nonetheless , Traditional privacy – preserving data stream mining environment which requires dynamic updating . For example for a massive amount of income data, the execution efficiency of traditional methods can no longer respond to user demand . Furthersome , the potential infinite number of data streams plus limited memory space has constrained the traditional methods from obtaining the mining result with accuracy . In view of the above mentioned issues , studies of privacy preservation in data stream mining in recent years has become one of the important issues in the field of data mining[18] .

Recently data streams are emerging as a new type of data, which are different from traditional static data . The characteristics of data streams are as follows[3] :

- Data has timing preference .
- Data distribution changes constantly with time .
- Amount of data is enormous .
- Data flows in and out with fast speed .

Traditional data mining algorithms are not designed for the static databases . If the data changes it is necessary ti rescan the database again , which leads to long computation time and inability to prompt to the user request[3] .

There exists different privacy preservation techniques as discussed in [14] , they proposed a random matrix-based spectral filtering technique to recover the original data from the perturbed data . They proposed two data reconstruction methods that are based on data correlations . One method uses the PCA – Principal Component Analysis and the other method uses Bayes estimate technique . Their study shows that when the co-relation between the data attributes is high , the original data can be reconstructed easily and more accurately i.e more privacy breaches will happen[14] .

The main idea of perturbation technique involves to perturb the original data by adding noise to the original data to preserve the sensitive information available in the data .

The objective of data mining is to generalize across population , rather than reveal information about individuals .But the main problem is that data mining works by evaluating individuals data[16] .

A. Need of privacy in data mining:

Every day we are leaving dozens of electronic trails through various activities such as using credit cards , swapping security cards , talking over phones and using emails . Ideally the data should be collected with the consent of the data subjects . The collector should provide some assurance to the individuals that the privacy will be protected .

II. PRIVACY PRESERVING IN DATA MINING

Due to the enormous benefits of data mining , yet high public concerns regarding the individual privacy , the implementation of privacy preserving data mining techniques have become a demand at the moment . A privacy preserving data mining technique provides the individual privacy while allowing the \extraction of the useful information[7] .

There are several different methods that can be used to enable privacy preserving data mining . One particular class of such techniques modifies the data set before its release , in an attempt to protect the individual records from being re – identified . An intruder with supplementary knowledge , can not be certain about the correctness of re – identification , when data set has been modified[3] .

High data quality and privacy are two important requirements that a good privacy preserving technique should satisfy . We need to evaluate data quality and the degree of privacy of a perturbed data set[3] .

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[18] . 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 sampling , local suppression , random noise rounding and micro – aggregation[5] .

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[5] . 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 sone data are suppressed and some details are removed.

The below figure – 1 shows the classification of different masking techniques :

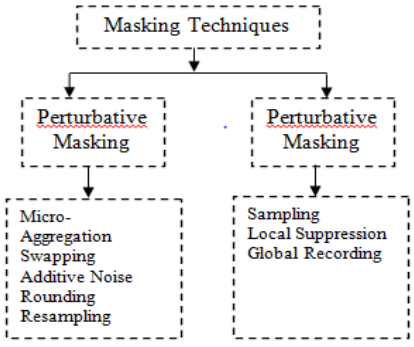


Fig. 1: Classification of Masking Techniques

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[19] 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 perturbated data[19] .

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 it's simple concept , easy implementation and comprehensible mining result[18]

C. Normalization Techniques for Privacy Preserving Data Mining

In[17] they have described the use of different normalization techniques 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 in[17] 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 $[new\_min_A , new\_max_A]$, 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 :

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

where Mean(P) = Sum f 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 :

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

Where m is the smallest integer that max(|d'|)< 1 .

IV. Evaluating Privacy Preserving Algorithms

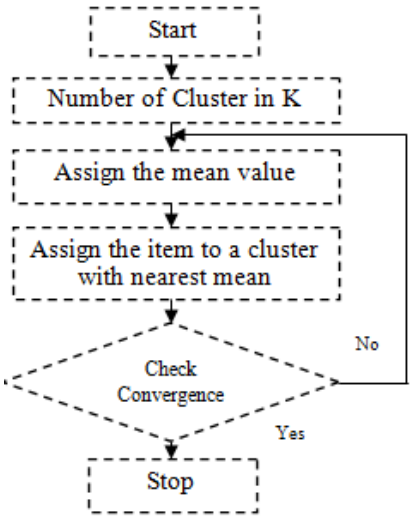
Another important aspect of privacy preserving data mining algorithms is their evaluation against certain parameters like[16] :

- Performance : The performance of a mining algorithm is measured in terms of the time required to achieve the privacy criteria .
- Data Utility : It is basically a measure of information loss or loss in the functionality of data in providing the results , which could be generated in the absence of PPDM algorithms .
- Uncertainty Level : It is a measure of uncertainty with which the sensitive information that has been hidden can still be predicted .
- Resistance : Resistance is a measure of tolerance shown by the PPDM algorithm against various data mining algorithms and models .

The main two important criteria are quantification of privacy and information loss . Quantification of privacy is a measure that indicates how closely the original value of an attribute can be estimated[16] . If it can be estimated with higher confidence , the privacy is low and vice – versa .

A. K – Means Clustering Algorithm :

Clustering is an un-supervised learning technique which groups the similar objects into appropriate clusters . Flowchart in figure – 2 summarizes the steps involved in K – means clustering algorithm.



V. Conclusion

Now a days, privacy is the most important approach to protect the sensitive data. People are very much worried about their sensitive information which they don't want to share. Our survey in this topic focuses on the existing techniques which have been already present in the field of Privacy Preserving Data Mining. From our analysis, we have found that there is no single technique that is used in all domains.

All methods perform in a different way depending on the type of data as well as the type of application or domain. But still from our analysis, we can conclude that Cryptography and Random Data Perturbation methods perform better than the other existing methods. Cryptography is best technique for encryption of sensitive data.

On the other hand Data Perturbation will help to preserve data and hence sensitivity is maintained. And at last, we want to say that perturbation technique with normalization is used to improve the level of privacy so perturbation technique with normalization is more important than all other existing techniques.

References

[1] Umamaheswari. K and S. Niraimathi. "A Study on Student Data Analysis Using Data Mining Techniques", International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), Vol. 3, Issue 3, pp.117-120, August 2013.

[2] M. Kholghi and M. Keyvanpour "An Analytical Framework For Data Stream Mining Techniques Based On Challenges And Requirements", International Journal of Engineering Science and Technology (IJEST), Vol. 3, No. 3, pp.2507-2513 Mar 2011.

[3] T.J. Trambadiya,and P. bhanodia "A Heuristic Approach to Preserve Privacy in Stream Data with Classification", International Journal of Engineering Research and Applications (IJERA), Vol. 3, Issue 1, pp.1096-1103, Jan -Feb 2013.

[4] V.S. Verykios, L.P. Provenza "State-of-the-art in Privacy Preserving Data Mining", SIGMOD Record, Vol. 33, No. 1, pp.50-57 March 2004.

[5] Dr.A.Tamilarasi and S. Vijayarani "A New Technique For Protecting Sensitive Data And Evaluating Clustering Performance" International Journal of Information Technology Convergence and Services (IJITCS), Vol.1, No.2, pp.7-18 April 2011.

[6] B. Pandya, U.K. Singh," An Overview of Traditional Multiplicative Data Perturbation", International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), Vol. 2, Issue 3, pp.424-429, March 2012.

[7] A.V Mary, Dr.T. Jebarajan "A Novel Data Perturbation Technique with Higher Security" international journal of computer engineering and Technology (IJCET), Vol. 3, Issue 2, pp.126-132, July- September (2012).

[8] D. Patil , T.N Rashmi, and S.M Akhtar, "Perturbation Based Reliability And Maintaining Authentication In Data Mining" International Conference on Advances in Computer and Electrical Engineering (ICACEE), pp.59-63, Nov 2012.
[9] K.Patel "Privacy-Preserving Data Stream Classification: An approach using MOA framework", GIT-Journal of Engineering and Technology, Vol. 6, 2013.
[10] N. Gupta, and I. Rajput, "Preserving Privacy Using Data Perturbation in Data Stream", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Vol. 2, No 5, May 2013.
[11] Yogita, D. Toshniwal "Clustering Techniques for Streaming Data-A Survey", 3rd IEEE International Advance Computing Conference (IACC) , pp.951-956, 2012
[12] K. Wankhade , T.Hasan and R.thool "A Survey: Approaches for Handling Evolving Data Streams", International Conference on Communication Systems and Network Technologies IEEE, pp. 621-625, 2013
[13] S.Guha, A.Meyerson, N.Mishra and R.Motvani "Clustering Data Streams: Theory and Practice", IEEE transactions on knowledge and data engineering, Vol. 15, NO. 3, pp. 515- 528, MAY/JUNE 2003
[14] H. Chhinkaniwala and S.Garg "Tuple Value Based Multiplicative Data Perturbation Approach To Preserve Privacy In Data Stream Mining" International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.3, pp. 53-61, May 2013 .
[15] Madjid Kalilian and Norwati Mustapha “Data Stream Clustering : Challenges and Issues” Proceedings of the International MultiConference of Engineering and Computer Scientists 2010 Vol1.
[16] Majid B. Malik , M. Asger Ghazi and Rashid Ali “ Privacy Preserving Data Mining Techniques : Cuurent Scenario and Future Prospects” . IEEE 2012 , PP : 26 - 32 .
[17] C.Saranaya and G.Manikandan “ A Study on normalization Techniques for Privacy Preserving Data Mining” . International journal of Engineering and Technology (IJIET) , Vol 5 No 3 Jun – Jul 2013 , PP : 2701 – 2704 .
[18] Syed md. Tarique Ahmad , Shameemul Haque & Prince Shoeb Khan “ Privacy Preserving in Dat Mining by Normalization ” . International journal of Computer Application (IJCA) , Vol 96 No 6 Jun 2014 , PP : 14 - 18 .
[19] Kumari P. , Raju K . “Privacy Preserving in Clustering using Fuzzy Sets ” . Proceedings of the 2006 International Conferenceon Data Mining , Las vegas , Nevada , USA 2006 , PP : 26 – 29

