ISSN: 2321-7782 (Online)

Volume 3, Issue 1, January 2015

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

Overview of Privacy preserving techniques and data accuracy

Brijal H. Patel[1]
Research Scholar, CSE Department
Parul Institute of Technology
Vadodara, India

Ankur N. Shah[2]
Assistant Professor, CSE Department
Parul Institute of Technology
Vadodara, India

Abstract: Privacy preserving allows sharing of privacy sensitive data for analysis purposes so it is very popular technique. so, people have ready to share their data. In recent years, privacy preserving data mining is an important one because wide availability of data is there. It is used for protecting the privacy of the critical and sensitive data and obtains more accurate results of data mining. The random noise is added to the original data in privacy preserving data Mining (PPDM) approach, which is used to publish the accurate information about original data. The main objective of privacy preserving data mining is to develop algorithms for modifying the original data and securing the information to be misused, so that the private data and private knowledge remain as it is after mining process. This topic is used to reiterate several privacy preserving data mining technologies to protect sensitive information's privacy and obtaining data clustering with minimum information loss for multiplicative attributes in dataset.

Keywords: Privacy, K-means clustering, precision, recall, DCT.

I. INTRODUCTION

Recent developments in information technology have enabled collection and processing of vast amounts of personal data, such as criminal records, shopping habits, credit and medical history, and driving records. This information is undoubtedly very useful in many areas, including medical research, law enforcement and national security. However, there is an increasing public concern about the individuals' privacy. Privacy is commonly seen as the right of individuals to control information about them.

Generally when we talk about privacy as "keep information about me from being available to others". Our real concern is that our information not be misused. The fear is there of information to be misuse once it is released. Utilizing this distinction ensuring that a data mining project won't enable misuse of personal information – opens opportunities that "complete privacy" would prevent. so we need technical and social solutions that ensure data will not be released. second view is corporate privacy the release of information about a collection of data rather than an individual data item. here concerns is not about the individual's personal information; but knowing all of them enables identity theft. This collected information problem scales to large, multi-individual collections as well.

Advantages of Privacy Protection are personal information protection, proprietary or sensitive information protection, enables collaboration between different data owners (since they may be more willing or able to collaborate if they need not reveal their information), and compliance with legislative policies. Concerns about informational privacy generally relate to the manner in which personal information is collected, used and disclosed. When a business collects information without the knowledge or consent of the individual to whom the information relates, or uses that information in ways that are not known to the individual, or discloses the information without the consent of the individual, informational privacy may be violated.

Achieving the privacy of the data in data stream model is quite difficult than traditional data mining model because of the characteristics of data stream model. There is continuous flow of the data and real time processing of the data object, incorporating privacy preserving phenomenon before making data available to classification or clustering algorithm requires interrupting rate of flow incoming data, doing some processing to achieve privacy. So algorithms should be appropriately configured to cope with the interruption. Because the whole data set are not available at the same moment, achieving privacy preserving in data stream model is more difficult.

II. Survey of Privacy Preserving Techniques

The concept of K-anonymity is discussed in [1]. The *k*1-anonymity privacy requirement, which requires each record in an anonymized table to be indistinguishable with at least *k*-other records within the dataset, with respect to a set of quasi-identifier attributes. To achieve the *k*-anonymity requirement, they used both generalization and suppression for data anonymization.

In general, *k*anonymity guarantees that an individual can be associated with his real tuple with a probability at most 1/*k*. In general, *k* anonymity guarantees that an individual can be associated with his real tuple with a probability at most 1/*k*. While *k*-anonymity protects against identity disclosure, it does not provide sufficient protection against attribute disclosure. There are two attacks: the homogeneity attack and the background knowledge attack. The other technique discussed in [1] is the perturbation approach, The perturbation approach works under the need that the data service is not allowed to learn or recover precise records. In the perturbation approach, the distribution of each data dimension reconstructed independently. This means that any distribution based data mining algorithm works under an implicit assumption to treat each dimension independently. Another branch of privacy preserving data mining which using cryptographic techniques was developed[1]. This technique is hugely popular for two main reasons: First, cryptography offers a well-defined model for privacy, which includes methodologies for proving and quantifying it. Second, there exists a vast toolset of cryptographic algorithms and constructs to implement privacy-preserving data mining algorithms.But, recent work has pointed that cryptography prevents privacy leaks in the process of computation. Thus, it falls short of providing a complete answer to the problem of privacy preserving data mining.

The another technique in [1] is randomized response techniques. The basic idea of randomized response is to scramble the data in such a way that the central place cannot tell with probabilities better than a pre-defined threshold whether the data from a customer contain truthful information or false information. The last technique in [1] is the condensation approach, which constructs constrained clusters in the data set, and then generates pseudo-data from the statistics of these clusters. This approach uses a methodology which condenses the data into multiple groups of predefined size, For each group, certain statistics are maintained. Each group has a size at least k, which is referred to as the level of that privacy-preserving approach. The greater the level, the greater the amount of privacy. At the same time, a greater amount of information is lost because of the condensation of a larger number of records into a single statistical group entity.

In [2] paper , the above discussed techniques are given and one more technique is discussed that is soft computing techniques. Soft computing is a consortium of methodologies which work synergistically and provides in one form or another flexible information processing capabilities for handling real-life ambiguous situations.

Soft computing techniques[2] include fuzzy logic, neural networks, genetic algorithms, and rough sets. Fuzzy sets provide a natural framework for the process in dealing with uncertainty. It makes it possible to model imprecise and qualitative knowledge as well as the transmission and handling of uncertainty at various stages. Neural Networks are widely used for classification and rule generation. Genetic algorithms are adaptive, robust, efficient and global search methods, suitable in situations where the search space is large. Rough set is a mathematical tool for managing uncertainty that arises from indiscernibility between objects in a set. In paper [5] all this techniques are repeated with some logical modifications. The *k*-anonymity[5] privacy requirement, which requires each record in an anonymized table to be indistinguishable with at least *k*-1 other records within the dataset, with respect to a set of quasi-identifier attributes. To achieve the *k*-anonymity requirement, they used both generalization and suppression for data anonymization. Unlike traditional privacy protection techniques such as data swapping and adding noise, information in a *k*-anonymous table through generalization and suppression remains truthful.

In the perturbation approach [5], the distribution of each data dimension is reconstructed independently. This means that any distribution based data mining algorithm works under an implicit assumption to treat each dimension independently. In many cases, a lot of relevant information for data mining algorithms such as classification is hidden in inter-attribute correlations

Another branch of privacy preserving data mining which using cryptographic techniques [5] was developed. This branch became hugely popular for two main reasons: Firstly, cryptography offers a well-defined model for privacy, which includes methodologies for proving and quantifying it. Secondly, there exists a vast toolset of cryptographic algorithms and constructs to implement privacy-preserving data mining algorithms.

Another two approach are discussed that is randomized response techniques and condensation approach[5]. A number of methods have recently been proposed for privacy preserving data mining of multidimensional data records. This paper intends to reiterate several privacy preserving data mining technologies clearly and then proceeds to analyze the merits and shortcomings of these technologies.

III. Data Stream Classification Using Different Approach

The methods that are based on data stream are given in [3]. The characteristics of data streams are: Data has timing preference; data distribution changes constantly with time; the amount of data is enormous; Data flows in and out with fast speed; and immediate response is required.

Proposed methods for data perturbation in [3]

1. This option is only for numeric values

Add the any value in attribute’s values (input from the user) suppose any attribute have value 12, 14, 11, 15, 9 etc. user give input 5, so add 5 to each value and output will be 17, 19, 16, 20, 14 etc.

2. This option is only for non-numeric values Change the non-numeric value of selected attribute by any other non-numeric value. (Suppose values is car1 so replace by selected value suppose p1 or other) (Used ASCII in programming)

1. Select non numeric attribute.
2. Find distinct values of selected attribute.
3. Generate distinct values mapping to each value of distinct values of selected attribute.
4. Replace old distinct values with generated/new distinct values.

3. This option is only for numeric and non-numeric values

Interchange the values of the same attribute (by randomly choose value only from that attribute)

1. Select Attribute from the data file.

2. Loop through all instances, I=0 To numberOfInstance

a. Randomly select instance/row and get Value of selected attribute of that instance/row.

b. Set randomly selected value to selected attribute of instance 'I'.

4. This option is only for numeric values

Find mean of numeric value of any particular row’s numeric attribute and replace chosen attribute value by this answer.

1. Select non numeric attribute.

2. Loop through all instances.

a. Find mean of numeric attribute of all each instance/row.

b. Set mean to selected attribute.

Data perturbation algorithm is given in [4]. Data perturbation refers to a data transformation process typically performed by the data owners before publishing their data. Two steps: date streams preprocessing and data streams mining. In the step of data streams preprocessing,the algorithm for data perturbation that is used for perturb the data using window approach algorithm. Perturbation techniques are often evaluated with two basic metrics: level of privacy guarantee and level of model-specific data utility preserved, which is often measured by the loss of accuracy for data classification. By using data perturbation algorithm they generate different perturbed dataset. And in the second step,apply the Hoeffding tree algorithm on perturbed dataset.The classification result of perturb dataset shows minimal information loss from original dataset classification[4].

Data magnets are techniques and tools used to collect personal data[6]. Examples of data magnets include explicitly collecting information through on-line registration, identifying users through IP addresses, software downloads that require registration, and indirectly collecting information for secondary usage. In many cases, users may or may not be aware that information is being collected or do not know how that information is collected.

In Paper[6] Principal Component Analysis (PCA) is used for transforming the multidimensional data into lower dimensions. PCA is a standard tool in modern data analysis. PCA assumes that all the variability in a process should be used in the analysis therefore it becomes difficult to distinguish the important variable from the less important A data set $\mathbf{x}i$, (i =1,...,n) is summarized as a linear combination of orthonormal vectors (called principal components):

$$\boldsymbol{f}\,(\mathbf{x},\mathbf{V}) = \mathbf{u} + (\mathbf{x}\mathbf{V})\mathbf{V}T$$

where $\boldsymbol{f}\,(\mathbf{x},\mathbf{V})$ is a vector valued function, $\mathbf{u}$ is the mean of the data $\mathbf{x}i$ { }, and $\mathbf{V}$ is an $d \times m$matrix with orthonormal columns. The mapping $\mathbf{z}i = \mathbf{x}i\mathbf{V}$ provides a low-dimensional projection of the vectors $\mathbf{x}i$ if $m < d$.

The first principal component is an axis in the direction of maximum variance. Principal components have the following optimal properties in the class of linear functions f ($\mathbf{x}$,$\mathbf{V}$): The principal components $\mathbf{Z}$ provide a linear approximation that represents the maximum variance of the original data in a low-dimensional projection. They also provide the best low-dimensional linear representation in the sense that the total sum of squared distances from data points to their projections in the space is minimized.If the mapping functions F and G are restricted to the class of linear functions, the composition $F(G(\mathbf{x}))$ provides the best (i.e., minimum empirical risk) approximation to the data. PCA is most appropriate for normal / elliptical distributions (where linear PCA approach provides the best possible solution).

IV. Tuple Value Based Multiplicative Data Perturbation Approach

Tuple Value Based Multiplicative Data Perturbation: To protect the sensitive attribute value, tuple value of instance to be processed is computed first. Tuple value is the average of normalized values (computed using -*score* normalization) of attributes of given instance except the class attribute.

The tuple values are then multiply with the values of sensitive attribute of respective instances. The resultant dataset with perturbed sensitive attribute values is likely preserves statistical characteristics of original dataset.

Precision and Recall are two important measures to determine the effectiveness and accuracy of the information retrieval system. Results of proposed approach have been quantified using precision and recall measures provided with MOA framework Accuracy using these two measures.

V. Enhance Technique of Perturbation Approach

Data perturbation refers to a data transformation process typically performed by the data owners before publishing their data. The goal of performing such data transformation is two-fold. On one hand, the data owners want to change the data in a certain way in order to disguise the sensitive information contained in the published datasets, and on the other hand, the data owners want the transformation to best preserve those domain-specific data properties that are critical for building meaningful data mining models, thus maintaining mining task specific data utility of the published datasets.

The stage of data streams pre-processing uses perturbation algorithm to perturb confidential data. Users can flexibly adjust the data attributes to be perturbed according to the security need. Therefore, threats and risks from releasing data can be effectively reduced.

Discrete cosine transformation is a unitary transformation, which can preserve the Euclidian distances between original and transformed domains. The focus of DCT is preserving Euclidian distances as well as privacy preservation. In DCT, data characteristics remain unchanged during whole process and behavior of data as same as original data.

VI. Conclusion

Privacy is the most important approach to protect the sensitive data. The data which they don't want to share ,People are very much worried about their sensitive information. My survey in this topic focuses on the existing techniques which have been already present in the field of Privacy Preserving Data Mining. From my analysis, I have found that, no single technique that is used in all domains. All techniques perform in a different way as the type of data as well as the type of application or domain. But still from my analysis, I can conclude that Random Data Perturbation and Cryptography techniques perform better than the other existing methods. Cryptography is best technique for encrypting sensitive information. On the other hand Data Perturbation will help to preserve data so sensitive information is maintained. And at last, I want to say that perturbation technique with normalization is used to improve the level of privacy with multiplicative attributes in dataset so perturbation technique with normalization is more important than all other existing techniques. Mainly, based on normalization value based perturbation algorithm is implemented and extended with sliding window concept. I will take clustering result and compare with result of single perturbation technique. Comparison based on same clustering algorithm in MOA framework for multiplicative attributes. Performance evaluation by k-means clustering method and calculating Accuracy based on available results.

Acknowledgment

I thanks to Assistant Professor Mr. Ankur N. Shah who is from Computer & Science Department of Parul Institute And Technology .he is also my guide who inspire me for doing such work and due to their support here I am able to write this review paper and she also provide me good knowledge about perturbation techniques and Guide me for doing this work.

References

1. W.T. Chembian1, Dr. J.Janet, "A Survey on Privacy Preserving Data Mining Approaches and Techniques", Proceedings of the Int. Conf. on Information Science and Applications ICISA 2010 Chennai, India.
2. Majid Bashir Malik, M. Asger Ghazi, Rashid Ali, "Privacy Preserving Data Mining Techniques: Current Scenario and future prospects ", 2012 Third International Conference on Computer and Communication Technology
3. Kiran Patel, Hitesh Patel, Parin Patel, "Privacy Preserving in Data stream classification using different proposed Perturbation Methods", © 2014 IJEDR | Volume 2, Issue 2 | ISSN: 2321-9939
4. 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.
5. Jian Wang ,Yongcheng Luo, Yan Zhao Jiajin Le," A Survey on Privacy Preserving Data Mining", First International Workshop on Database Technology and Applications, 978-0-7695-3604-0/09 © 2009 IEEE
6. R.Vidya Banu, N.Nagaveni," Preservation of Data Privacy using PCA based Transformation", International Conference on Advances in Recent Technologies in Communication and Computing, © 2009 IEEE
7. Hitesh Chhinkaniwala1 and Sanjay 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, May 2013
8. "DATA MINING" concepts and techniques by Jiawei Han, Michleline Kamber and Jian pei Third Edition . ELSEVIER , Morgan Kaufmann publisher

Author(s) Profile

Brijal H. Patel received the B.E. degree in Information Technology from Sigma Institute of Engineering, Gujarat Technological University in 2013.

Currently she is doing her M.E. Degree from Parul Institute of Technology of Gujarat Technological University and her interesting area of research is in Privacy Preserving of Data Mining.