Scientific Journal of Impact Factor (SJIF): 4.14

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 3, Issue 2, February -2016

A Review on Privacy Preserving Data Mining Using Piecewise Vector Quantization

Chauhan Harshil Narendrabhai [1],Lolita Singh[2]

[1]*Computer Engineering Department, HJD Institute of Technical Education and Research, Kera-Kutch,*
[2] *Computer Engineering Department, HJD Institute of Technical Education and Research, Kera-Kutch,*

Abstract *— Now in days, Privacy Preserving Data Mining (PPDM) is the most important area of the Data Mining, which aims to protect the sensitive data comes from different businesses and organizations. Such data is very useful for data owner to make decision based on the mining result. But the privacy of the data may prevent the data owners from sharing their data for analysis purpose. The Privacy Preserving Data Mining has become increasingly popular because it allows sharing sensitive data for analysis purposes. In this paper we present some key directions in the field of privacy preserving data mining and various dimensions of privacy preserving techniques. We provide a review on piecewise vector quantization approach which segmentized each row of dataset and quantization approach will performed on each segment using K-means which later are united to form a transformed dataset.*

Keywords- *Privacy Preserving Data Mining, Clustering, Piecewise Vector Quantization, Codebook, K-means*

I. INTRODUCTION

Data mining is a subfield of the computer science. Data mining means to extract the hidden information from the large database. It's also referred as a knowledge Discovery from Data, or KDD. The goal of data mining is the extraction of patterns and knowledge from large dataset, not the extraction (mining) of the data itself [12]. Data mining can be very useful to predict the future trends and patterns, and also allowing the businesses and organizations to make the decisions based on that data mining result. The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown, interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection), and dependencies (association rule mining) [12].

This paper provides a review on the various techniques which used to preserve the privacy of the sensitive data. In this we provide the survey on the piecewise vector quantization, in which the K-means clustering algorithm can used to cluster the segmentized data according to their similarities, after that the data can be transformed into another form by quantization process. At the end accuracy between original data and transformed can be measured. The key point of this approach is there is no data compression but there is quantization of data so that privacy is preserved [1].

II. PRIVACY PRESERVING DATA MINING

Privacy preserving data mining (PPDM) is one of the most important areas of the data mining that aims to protect the sensitive data and do not reveal the personal identity. Over the last twenty years, the growth of storing private data of individuals was increased. Those data are very important for predict the useful patterns and future trends. But the easy access of the personal data may restrict the owner to share their data. The main purpose of the Privacy preserving data mining (PPDM) is to develop algorithms for transforming the original data in another form, so that the private data and knowledge stay protected even after the data mining process.

Privacy preserving data mining (PPDM) algorithm must hold the following two restrictions:

- When transformation applied to original dataset must preserve the privacy of individuals. So that modified dataset conceals the values of confidential attributes.
- The similarity between objects in modified dataset must be the same as that one in original dataset, or just slightly altered by the transformation process.

The key directions in the field of privacy-preserving data mining are as follows [1] [5]:

1. Privacy-Preserving Data Publishing

These techniques tend to study different transformation methods associated with privacy. These techniques include methods such as randomization, Anonymization, and *l*-diversity. Another related issue is how the perturbed data can be used in conjunction with classical data mining methods such as association rule mining. Other related problems include that of determining privacy-preserving methods to keep the underlying data useful (utility-based methods), or the problem of studying the different definitions of privacy, and how they compare in terms of effectiveness in different scenarios.

2. Changing the results of Data Mining Applications to preserve privacy

In many cases, the results of data mining applications such as association rule or classification rule mining can compromise the privacy of the data. This has spawned a field of privacy in which the results of data mining algorithms such as association rule mining are modified in order to preserve the privacy of the data. A classic example of such techniques are association rule hiding methods, in which some of the association rules are suppressed in order to preserve privacy.

3. Query Auditing

Such methods are akin to the previous case of modifying the results of data mining algorithms. Here, we are either modifying or restricting the results of queries.

4. Cryptographic Methods for Distributed Privacy

In many cases, the data may be distributed across multiple sites, and the owners of the data across these different sites may wish to compute a common function. In such cases, a variety of cryptographic protocols may be used in order to communicate among the different sites, so that secure function computation is possible without revealing sensitive information.

5. Theoretical Challenges in High Dimensionality

Real data sets are usually extremely high dimensional, and this makes the process of privacy-preservation extremely difficult both from a computational and effectiveness point of view. It has been shown that optimal k-anonymization is NP-hard. Furthermore, the technique is not even effective with increasing dimensionality, since the data can typically be combined with either public or background information to reveal the identity of the underlying record owners.

III. CLASSIFICATION OF PRIVACY PRESERVING TECHNIQUES

There are many methods available for privacy preserving data mining. Privacy preserving data mining techniques can be classified based on the following dimensions, as in [2] [8].

3.1 Data distribution
3.2 Data modification
3.3 Data mining algorithm
3.4 Data or rule hiding
3.5 Privacy preservation

3.1 Data Distribution

The first dimension is distribution of the data. This dimension also classified as horizontal data distribution and vertical data distribution [3].

Horizontal data distribution divides dataset into non-overlapping horizontal partition. In this different places have different records about same entities. In vertical data distribution each sites have different number of attributes with same number of transaction.

3.2 Data Modification

The data modification is used in order to change the original values that need to be changed to preserve the privacy of the individuals. The data modification has following methods:

- Perturbation: can be done by altering the original values with new values or by adding noise.
- Blocking: which is replacement of the existing attribute value with an aggregation?
- Swapping: refers to interchanging the values of the individual records.
- Sampling: refers to losing data for only sample of a population.

3.3 Data Mining Algorithm

In this dimension data modification is also present here. In this various data mining algorithms are used for hiding the original data. The data mining algorithm like Association rule mining, Clustering algorithm, Bayesian networks, Classification are used.

3.4 Data or Rule Hiding

In this dimension, the aggregated data can be hidden in form of rules because its complexity is higher, that's why developer refers this most. It's also known as 'Rule Confusion' [3].

3.5 Privacy Preserving

The privacy preserving dimension is the most important technique used for selective data modification. This dimension mainly focuses on the modification of the data in such a manner that after modification the original data not loss. The Privacy preserving dimension techniques are:

3.5.1 Randomization
3.5.2 Anonymization
3.5.3 Encryption Method

3.5.1 Randomization

In privacy preserving, the randomization is an efficient and inexpensive method. In randomization some additional data can be added in the original data. Noise can be introduced by adding or multiplying random values to numerical records or by deleting real item and adding fake values to the set of attributes. The randomization method can be easily implemented at *data collection time*, because the noise added to a given record is independent of the behavior of other data records. This is also a weakness because outlier records can often be difficult to mask [5].

3.5.2 Anonymization

The Anonymization was developed because of the possibility of the indirect identification of data from pubic datasets. The k-anonymity is an attractive technique because of the simplicity of the definition and the numerous algorithms available to perform the Anonymization [5]. Suppression and Generalization are two methods which commonly used to achieve k-anonymity for some value of k.

Suppression consists in protecting sensitive information by removing it. Suppression, which can be applied at the level of single cell, entire tuple, or entire column, allows reducing the amount of generalization to be enforced to achieve k-anonymity. If a limited number of outliers would force a large amount of generalization, then such outliers can be removed thus allowing satisfaction of k-anonymity with less generalization [5].

In generalization method, individual values of attributes are replaced with a generalized version of them. For example, the age of the person could be generalized to a range such as youth, middle age, and adult without specifying appropriately [7].

3.5.3 Encryption Method

Encryption is a well-known technique for preserving privacy of sensitive data. It resolves the problem that people jointly conduct mining tasks based on some private inputs they provide. These mining tasks occur between two competitors or between untrusted parties. Encryption method ensures that the data transfer is secure and exact. Encryption method guarantees very high level of data privacy.

IV. PIECEWISE VECTOR QUANTIZATION

4.1 Vector Quantization

Vector Quantization is mainly used in signal compression and coding. It is a lossy compression method based on principle block coding [1]. It works by dividing a large set of points (vectors) into groups having approximately the same number of points closest to them. Each group is represented by its centroid point, as in k-means and some other clustering algorithms [13]. It is used in privacy preserving by approximating each point (row of data) to the other with the help of vector quantization approach (VQ).There is no data compression but there is data quantization, so that privacy is preserved [1].

Vector Quantization process has following steps as given in [2] [9]:

1. Original Data
2. Constructing Codebook
3. Encoding Original data with Codebook and
4. Decoding

In vector quantization, a dataset will be taken. The codebook is generated from these training vectors using clustering techniques. After that in decoding process, an original data is divided into k-dimension vectors and each vector encoded by the index of codeword. In decoding process, encoded data can be reconstructed into original data using the same codebook.

4.2 Piecewise Vector Quantization Approach

In Piecewise Vector Quantization following steps are involved [1]:

Step 1: Input:

Input is dataset which is stored in file which contains sensitive information which is to be preserved such that there is less information loss and hence good clustering result. Each point (row) of data is sequence of real value X = x1 x2 x3…….xn. Dataset contain "m" row of data.

Step 2: Segmentation:

Dataset is segmentized into w datasets by decomposing each row of data into w segments each of length L. Dataset D is decomposed into D1 D2 D3….DW dataset, where each row of each dataset is of length L. If total number of attributes in original is not perfectly divisible by L then extra attributes is added with zero value which does not affect the result and later it will be removed at step 5.

Step 3: Clustering for Codebook Generation:

In order to generate codebook which is helpful in data transformation, K means clustering algorithm is used. By using K-means (or clustering) we can group the items into k clusters such that all items in same cluster are as similar to each other as possible. And items not in same cluster are as different as possible. We use the distance measures to calculate similarity and dissimilarity.

Step 4: Data Transformation by Quantization:

Each decomposed dataset Di is transformed into new dataset Di' by replacing each of the point (row data) with the point which fall nearest to it in its codebook. That is the point is replaced by the cluster centroid in which it falls.

Step 5: Reformation of Dataset:

Each segment Yi of row data X formed as segmentation step in step 2 is transformed into Zi by step 4. Now all the w transformed segment of each row is joined in the same sequence as segmentized in step 2 to form a new n dimensional transformed row data which replace the X in the original dataset.

Step 6: - Comparison for accuracy from distortion in data:

Clustering by K means is performed on original dataset and result received (R) Clustering by K means is performed on modified dataset and result received (R2) Comparison between the two result (R1 and R2) using Fmeasure metric and distortion measure.

IV. CONCLUSION

Privacy is the major concern to protect the sensitive data because of extensive growth in collection of sensitive data. All the existing methods perform in a different way depending on the type of data and also the type of application. From this study, we can conclude that the piecewise vector quantization is more effective and efficient method for privacy preservation. In piecewise vector quantization, the K-means algorithm can use centroid to cluster the data. But the K-means algorithm has limitation that it unable to handle the noisy data and also more sensitive to outliers. In Future work, we want to replace the k-means algorithm with other clustering algorithm like K-medoid algorithm that more robust than the K-means algorithm in presence of noisy and outliers.

REFERENCES

[1] S. Sasikala, S. Nathira Banu "Privacy Preserving Data Mining Using Piecewise Vector Quantization (PVQ)" Proceeding in IJARCST Vol. 2 Issue 3 (July-Sept 2014).

[2] D.Aruna Kumari, B. Pooja Y, A.Shalini B, Vinay "Privacy Preserving Data Mining Using LBG And ELBG", Proceedings of 6th IACEECE-2013, 29th September 2013, Chennai, India.

[3] Ms. Dhanalakshmi.M and Mrs.Siva Sankari.E "Privacy Preserving Data Mining Techniques-Survey" In proceeding IEEE International conference on Information Communication & Embedded System "ICICES 2014".

[4] Xueyun Li, Zheng Yan, and Peng Zhang "A Review on Privacy-Preserving Data Mining" IEEE International Conference on Computer and Information Technology-2014.

[5] Agarwal Charu C., Yu Philip S., Privacy Preserving Data Mining: Models and Algorithms, New York, Springer, 2008.

[6] Shalini S Singh, N C Chauhan "K-means v/s K-medoids: A Comparative Study" National Conference on Recent Trends in Engineering & Technology- MAY-2011.

[7] Tamanna Kachwala, Sweta Parmar "An Approach for Preserving Privacy in Data Mining" IJARCSSE Vol. 4, Issue 9, September 2014.

[8] D.Aruna Kumari, Dr.Rajasekhara Rao, and M.Suman "Privacy Preserving Clustering in Data Mining Using VQ Code Book Generation" CS&IT 2012

[9] D.Aruna Kumari, K. Rajasekhara Rao, M.Suman, Hima Bindu Y, Srividya B, Kamala G "Vector Quantization for Privacy Preserving Data Mining" Proceedings of International Academic Conference on Electrical, Electronics and Computer Engineering, 8th Sept. 2013, Chennai, India

[10] Tzu-Chuen Lu, Ching-Yun Chang "A Survey of VQ Codebook Generation" Journal of Information Hiding and Multimedia Signal Processing, Vol 1, Number 3, July 2010.

[11] D.Aruna Kumari, Dr.K.Rajasekhara Rao, M.Suman and Tharun Maddu "Compression in Privacy Preserving Data Mining" COMPUSOFT- An International Journal of advanced Computer Technology, Vol. 2, Issue 4, April-2013.

[12] Wikipedia. Data Mining: https://en.wikipedia.org/wiki/Data_mining.
[13] Wikipedia. Vector Quantization: https://en.wikipedia.org/wiki/Vector_quantization.
[14] Data Mining: Concepts and Techniques by Jiawei Han and Micheline Kamber 2nd Edition.

