ISSN:2320-0790

An International Journal of Advanced Computer Technology

Compression in Privacy preserving Data Mining

D.Aruna Kumari[1], Dr.K.Rajasekhara Rao[2], M.Suman[3] and Tharun Maddu[4]

[1,3] Department of Electronics and Computer Engineering, Associate professors ,CSI Life Member K.L.University, Vaddeswaram,Guntur
[1]enteraruna@yahoo.com, [3]suman.maloji@gmail.com

[2]Department of Computer Science and Engineering ,professor, K.L.University, Vaddeswaram,Guntur
[2]rajasekhar.kurra@klce.ac.in, [4]tharunmaddu@hotmail.com

[1,2,3]CSI LIFE MEMEBERS, [3]CSI-AP Student CO-coordinator

Abstract: Large Volumes of detailed personal data is regularly collected and analyzed by applications using data mining, sharing of these data is beneficial to the application users. On one hand it is an important asset to business organizations and governments for decision making at the same time analysing such data opens treats to privacy if not done properly. This paper aims to reveal the information by protecting sensitive data. We are using Vector quantization technique based on LBG Design algorithm for preserving privacy with the help of Codebook. Quantization will be performed on training data samples it will produce transformed data set. This transformed data set does not reveal the original data. Hence privacy is preserved

Keywords : Vector quantization, code book generation, privacy preserving data mining ,k-means clustering.

I. INTRODUCTION

Privacy preserving data mining (PPDM) is one of the important areas of data mining that aims to provide security for secret information from unsolicited or unsanctioned disclosure. Data mining techniques analyzes and predicts useful information. Analyzing such data may opens treat to privacy .The concept of privacy preserving data mining is primarily concerned with protecting secret data against unsolicited access. It is important because Now a day's Treat to privacy is becoming real since data mining techniques are able to predict high sensitive knowledge from huge volumes of data [1].

Authors Agrawal & Srikant introduced the problem of "privacy preserving data mining" and it was also introduced by Lindell & Pinkas. Those papers have concentrated on privacy preserving data mining using randomization and cryptographic techniques. Lindell and Pinkas designed new approach to PPDM using Cryptography but cryptography solution does not provides expected accuracy after mining result. And agrawal and srikanth focused on randomization and preserving privacy when the data is taken from multiple parties. When the data is coming from multiple sources then also privacy should be maintained. Now a days this privacy preserving data mining is becoming one of the focusing area because data mining predicts more valuable information that may be beneficial to the business, education systems, medical field, political ,…etc.

I. RELATED WORK

Many Data modification techniques are discussed in [1][3][4]

A. Perturbation or Randomization: One approach to privacy in data mining is to obscure or randomize data [2] making private data available by adding enough noise to it. In this case there is one server and multiple clients will operate ,Clients are supposed to send their data to server to mining purpose , in this approach each client adds some random noise before sending it to the server. So Sever will perform mining on that randomized data

B. Suppression

Another way of preserving the privacy is suppressing the sensitive data before any disclosure or before actual mining takes place. Generally Data contains several attributes, where some of the attributes may poses personal information and some of the attributes predicts valuable information. So we can suppress the attributes in particular fashion that reveals the personal information.

They are different types are there

1. Rounding
2. Generalization

In rounding process the values like 23.56 will be rounded to 23 and 25.77 rounded to 26…etc

In generalization process, values will be generalized like an address is represented with zip code.

If data mining requires full accesses to the entire database at that time all this privacy preserving data mining techniques are not required.

III. PROPOSED APPROACH

A. Privacy preserving clustering:

Privacy-preserving clustering aims to protect the secure attribute values of objects under clustering task. The problem of privacy preservation in clustering can be defined as follows as in [6][7]:

Let *D* be a relational database and *C* a set of clusters generated from *D*.

This paper aims to transform *D* to D' so that the following conditions are satisfied:

We Transform the Database D to D` By applying Vector quantization technique. We a Transformation is applied Database must preserve the privacy of individual, So that the transformed database D` closes the confidential attributes.

The similarity between the objects in *D' will* be the same as that one in *D*, or just minor alterations by the transformation process. The transformed database *D'* looks different from *D*, but the clusters in *D* and *D' could* be as close as possible, because we need to preserve data as well as should get accurate mining results.

In this paper, we are introducing LBG Algorithm for Data Transformation .Vector quantization is based on LBG algorithm is introduced in this paper. Quantization is a process of mapping infinite set of scalar or vector quantities to finite set of scalar or vector quantities. We can say that the Transformed dataset is compressed form of original dataset

B. Vector Quantization:

Quantization maps data from one form to another form. Generally used when compression is required. There are two types of quantization are there:

1. Scalar quantization
2. Vector quantization

Vector Quantization is one of the lossy data compression techniques; it maps infinite set of vector data to finite set of vector data.

There are several ways in which vector quantization has been implemented.
LBG (Linde, Buzo, and Gray)
ELBG (Enhanced Linde, Buzo, and Gray)
PCA (Principal Component Analysis)
Neural Network
Genetic Based Approach.
Generally Vector Quantization is used in signal processing but VQ can also be applied on multi-dimensional data. Ours is a relational data base. We can apply VQ for Our PPDM problem.

A VQ is just like an approximate [19][20]. It works based on the idea of ``rounding-off'' (compressing to the nearest integer). An example of a 1-dimensional VQ is shown below:

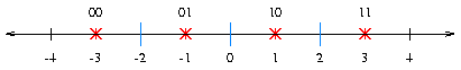


Fig 1: One Dimensional VQ

Here, every number less than -2 is approximated by -3. Numbers between -2 and 0 are approximated by -1. Every number between 0 and 2 are approximated by +1. Every number greater than 2 is approximated by +3. Approximate values are uniquely represented by 2 bits.

An example of a 2-dimensional VQ is shown below:

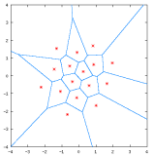


Fig 2: 2 Dimensional VQ

Here some regions have been given; regions are approximated by red color star. In the above diagram we have 16 regions so we have 16 red stars.

Here, every pair of numbers falling in a particular region is approximated by a red star associated with that region. Note that there are 16 regions and 16 red stars -- each of which can be uniquely represented by 4 bits. Thus, this is a 2-dimensional, 4-bit VQ. Its rate is also 2 bits/dimension.

In the above two examples, the red stars are called *codevectors* and the regions defined by the blue borders are called *encoding regions*. The set of all codevectors is called the *codebook* and the set of all encoding regions is called the *partition* of the space.

As stated in [7] the design of a Vector Quantization-based system mainly consists of three steps:

• Constructing a codebook from a set of training samples;
• Encoding the original signal with the indices of the nearest code vectors in the codebook;
• Using an index representation to reconstruct the signal by looking up in the codebook.

For our PPDM problem , reconstructing the original data is not required, so above two steps are involved such that it is difficult to get the original data back hence privacy is preserved.

C. Design Problem

The VQ design problem can be defined fallows. For a given relational database apply LGB algorithm that iteratively finds the regions and data that belongs to the specified region based on mean value. Based on mean value it will split the data into regions, again for every region mean will be calculated … this process is repeated until we get desired cluster objects that is called as codebook.

Let us assume that the training data consisting of M source vectors.

$$\mathcal{T} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M\}.$$

The above training data can be obtained from some huge database. Suppose, if the source is a speech signal, then the training data can be obtained from recording voice conversations.
M is assumed to be sufficiently large so that all the statistical properties of the source are captured by the training sequence. We assume that the source vectors are k-dimensional, e.g.

$$\mathbf{x}_m = (x_{m,1}, x_{m,2}, \ldots, x_{m,k}), \quad m = 1, 2, \ldots, M.$$

Let N be the number of code vectors and let

$$\mathcal{C} = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_N\},$$

represents the codebook. Each code vector is K- dimensional, e.g.,

$$\mathbf{c}_n = (c_{n,1}, c_{n,2}, \ldots, c_{n,k}), \quad n = 1, 2, \ldots, N.$$

Let S_n be the encoding region associated with code vector $\mathbf{c}_n$ and let

$$\mathcal{P} = \{S_1, S_2, \ldots, S_N\},$$

denote the partition of the space. If the source vector $\mathbf{x}_m$ is in the encoding region S_n, then its approximation (denoted by $Q(\mathbf{x}_m)$) is $\mathbf{c}_n$:

$$Q(\mathbf{x}_m) = \mathbf{c}_n, \quad \text{if } \mathbf{x}_m \in S_n.$$

E. Code Book Generation using LBG Design Algorithm

The LBG VQ design algorithm is an iterative algorithm which alternatively solves the above two optimality criteria. The algorithm requires an initial codebook C^0. [21][22][23]This initial codebook is obtained by the *splitting* method. In this method, an initial code vector is set as the average of the entire training sequence. This code vector is then split into two. The iterative algorithm is run with these two vectors as the initial codebook. The final two code vectors are spitted into four and the process is repeated until the desired number of code vectors is obtained. The algorithm is summarized below.

LBG Design Algorithm

LBG algorithm is used to generate codebook. Finally code book contains centroids.

1. Initially the codebook generation requires a Training sequence which is the input to LBG algorithm. The training sequence is

obtained from UCI Data repositories water plant treatment data set [23]

2. Let „R" be the region of the training sequence.

3. Generate an initial codebook from the training sequence, now it will be the centroid or mean of the training dataset and let the initial codebook be „". C

4. Split the initial code book in to C_n^- and C_n^+

   Where $C_n^+ = C(1+\in)$

   And $C_n^- = C(1-\in)$

   $\in$=0.01 is the minimum error to be obtained between old and new codewords.

5. Compute the difference between the training sequence and each of the codewords C_n^- and C_n^+ and let the difference be D^1.

6. Split the training sequence into two regions R1 and R2 depending on the difference „D" between the training sequence and the codewords C_n^- and C_n^+.

   The training vectors closer to C_n^+ falls in the region R1 and the training vectors closer to C_n^- falls in the region R2.

7. Let the training vectors falling in the region R1 be TV1 and the training sequence vectors falling in the region R2 be TV2.

8. Obtain the new centroid or mean for TV1 and TV2. Let the new centroids be CR1 and CR2.

9. Replace the old centroids C_n^+ and C_n^- by the new centroids CR1 and CR2

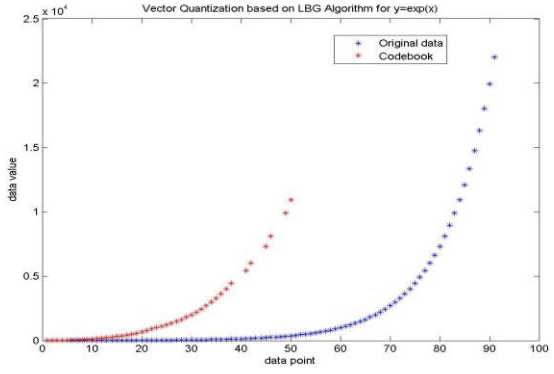
10. Compute the difference between the training sequence and the new centroids CR1 and CR2 and let the difference be D 1 .

11. Repeat steps 5 to 10 until $\frac{D^1 - D}{D} < \in$

12. Repeat steps 4 to 11 till the required number of codewords in the codebook are obtained. where $N=2^b$ represents the number of codewords in the codebook and „ b" represents the number of bits used for codebook generation, D represents the difference between the training sequence and the old codewords and D^1 represents the difference between the training sequence and the new codewords[23][27][28].

V. RESULTS

We have implemented above LBG algorithm using Matlab Software, and tested the results. In the output screen shots Blue line represents original data and red line represents Codebook that is compressed form of original data , hence it does not reveal the complete original information and it will reveal only cluster centroids



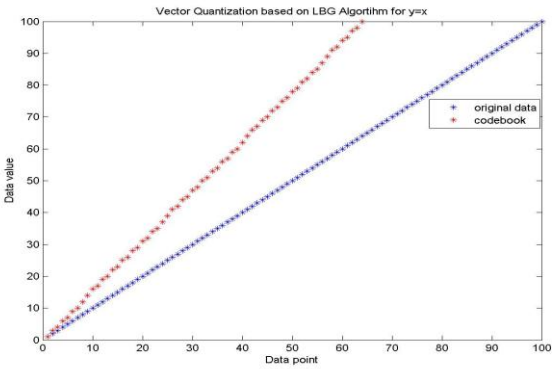


Figure 1: VQ based on LBG design Algorithm y=exp(x); Figure 2: VQ based on LBG design Algorithm y=x;

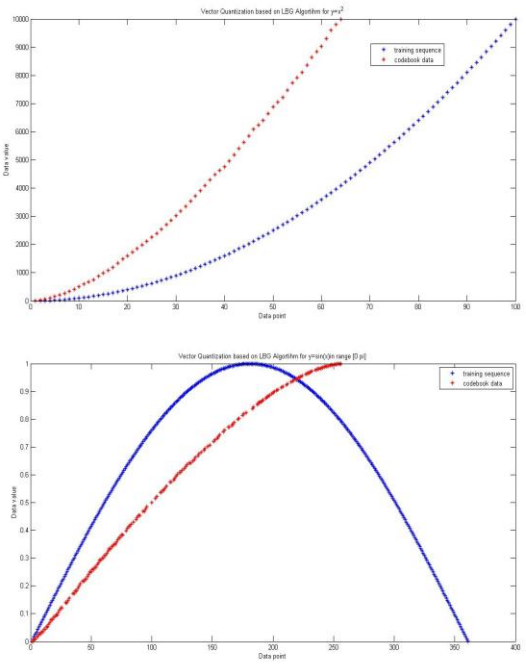


Figure 3: VQ based on LBG design Algorithm $y=x^2$;

Figure4: VQ based on LBG design Algorithm y=sin(x);

VI. CONCLUSIONS

This work is based on vector quantization, it is a new approach for privacy preserving data mining, upon applying this encoding procedure one cannot reveal the original data hence privacy is preserved. At the same time one can get the accurate clustering results. Finally we would like conclude that Efficiency depends on the code book generation.

VII. REFERENCES

[1] D.Aruna Kumari , Dr.K.Rajasekhar rao, M.suman " Privacy preserving distributed data mining using steganography "In Procc. Of CNSA-2010, **Springer Libyary**

[2] T.Anuradha, suman M,Aruna Kumari D "Data obscuration in privacy preserving data mining in Procc International conference on web sciences ICWS 2009.

[3] Agrawal, R. & Srikant, R. (2000). Privacy Preserving Data Mining. In Proc. of ACM SIGMOD Conference on Management of Data (SIGMOD'00), Dallas, TX.

[4] Alexandre Evfimievski, Tyrone Grandison Privacy Preserving Data Mining. IBM Almaden Research Center 650 Harry Road, San Jose, California 95120, USA

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

[6] Oliveira S.R.M, Zaiane Osmar R., A Privacy-Preserving Clustering Approach Toward Secure and Effective Data Analysis for Business Collaboration, In Proceedings of the International Workshop on Privacy and Security Aspects of Data Mining in conjunction with ICDM 2004, Brighton, UK, November 2004.

[7] Wang Qiang , Megalooikonomou, Vasileios, A dimensionality reduction technique for efficient time series similarity analysis, Inf. Syst. 33, 1 (Mar.2008), 115- 132.

[8] UCI Repository of machine learning databases, University of California, Irvine.http://archive.ics.uci.edu/ml/

[9] Wikipedia. Data mining. http://en.wikipedia.org/wiki/Data_mining

[10] Kurt Thearling, Information about data mining and analytic technologies http://www.thearling.com/

[11] Flavius L. Gorgônio and José Alfredo F. Costa "Privacy-Preserving Clustering on Distributed Databases: A Review and Some Contributions

[12] D.Aruna Kumari, Dr.K.rajasekhar rao,M.Suman "Privacy preserving distributed data mining: a new approach for detecting network traffic using steganography" in international journal of systems and technology(IJST) june 2011.

[13] Binit kumar Sinha "Privacy preserving clustering in data mining".

[14] C. W. Tsai, C. Y. Lee, M. C. Chiang, and C. S. Yang, A Fast VQ Codebook Generation Algorithm via Pattern Reduction, *Pattern Recognition Letters*, vol. 30, pp. 653{660, 2009}

[15] K.Somasundaram, S.Vimala, "A Novel Codebook Initialization Technique for Generalized Lloyd Algorithm using Cluster Density", International Journal on Computer Science and Engineering, Vol. 2, No. 5, pp. 1807-1809, 2010.

[16] K.Somasundaram, S.Vimala, "Codebook Generation for Vector Quantization with Edge Features", CiiT International Journal of Digital Image Processing, Vol. 2, No.7, pp. 194-198, 2010.

[17] Vassilios S. Verykios, Elisa Bertino, Igor Nai Fovino State-of-the-art in Privacy Preserving Data Mining in SIGMOD Record, Vol. 33, No. 1, March 2004.

[18] Maloji Suman,Habibulla Khan,M. Madhavi Latha,D. Aruna Kumari "Speech Enhancement and Recognition of Compressed Speech Signal in Noisy Reverberant Conditions " **Springer** - Advances in Intelligent and Soft Computing (AISC) Volume 132, 2012, pp 379-386

[19] M.Suman, K.B.N.Prasanna Kumar, K.Kavindra Kumar, G. Phrudhi Teja" A new approach on Compression of Speech Signals using MSVQ and its Enhancement Using Spectral Subtraction under Noise free and Noisy Environment" in Advances in Digital Multimedia ,Vol. 1, No. 1, March 2012,World Science Publisher, United States

[20] M. Suman, Habibulla khan, M. Madhavi Latha, D. Aruna kumari "Dimensions of performance in compressed speech signals and its enhancement", International journal of engineering sciences research(IJESR).

[21] M.Suman, K.B.N.P.Kumar, G.P.Teja,T.V.B hargava, 'Speech enhancement and recognition of compressed speech signal in noisy reverberant conditions', International research journal for signal processing.(IRJSP)

[22] M.Suman, M.satya sai ram, Dr. habibulla khan "Compression of Speech signals : LBG Algorithm "in Third international conference on SEEC 2010, cochin kerala

[23] M.Madhavi Latha, M.Satya Sai Ram and P.Siddaiah . Multi Switched Split Vector Quantizer, International journal of Computer, Information and systems science and engineering, Vol 2 No: 1

[24] M.Madhavi Latha, M.Satya Sai Ram and P.Siddaiah\Multi Switched Split Vector Quantization, Proceedings of World Academy of Science, Engineering and Technology Volume 27 Feb 2008 ISSN 1307-6884.