PRIVACY PRESERVING DATA MINING BASED ON VECTOR QUANTIZATION

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

[1,3]Department of Electronics and Computer Engineering, Associate professors ,CSI Life Member K.L.University, Vaddeswaram,Guntur
[1]Aruna_D@kluniversity.in and [3]suman.maloji@gmail.com
[2]Department of Computer Science and Engineering ,professor, K.L.University, Vaddeswaram,Guntur
[2]krr@kluniversity.in
[1,2,3]CSI LIFE MEMEBERS, [3]CSI-AP Student CO-coordinator

ABSTRACT

Huge Volumes of detailed personal data is continuously collected and analyzed by different types of applications using data mining, analysing such data is beneficial to the application users. It is an important asset to application users like business organizations, governments for taking effective decisions. But analysing such data opens treats to privacy if not done properly. This work aims to reveal the information by protecting sensitive data. Various methods including Randomization, k-anonymity and data hiding have been suggested for the same. In this work, a novel technique is suggested that makes use of LBG design algorithm to preserve the privacy of data along with compression of data. Quantization will be performed on training data it will produce transformed data set. It provides individual privacy while allowing extraction of useful knowledge from data, Hence privacy is preserved. Distortion measures are used to analyze the accuracy of transformed data.

KEYWORDS

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

1. INTRODUCTION

Privacy preserving data mining (PPDM) is one of the important area 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 days 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.

2. RELATED WORK

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

A. Perturbation or Randomization:

Agrawal and Srikant (2000) Introduced the randomization algorithm for PPDM, Randomization allows a several number of users to submit their sensitive data for effective centralized data mining while limiting the disclosure of sensitive values. It is relatively easy and effective technique for protecting sensitive electronic data from unauthorized use. In this case there is one server and multiple clients will operate ,Clients are supposed to send their data to server for 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 access to the entire database at that time all this privacy preserving data mining techniques are not required.

C. Cryptography

This is Also one of the famous approach for data modification techniques, Here Original Data will be encrypted and encrypted data will be given to data miners. If data owners require original data back they *will apply decryption techniques.*

3. VECTOR QUANTIZATION

This is the new technique proposed by (D.Aruna Kumari, Dr.k.Rajasekhara Rao, Suman,)), it transforms the original data to a new form using LBG. The design of a Vector Quantization consist of following steps:

- Design a codebook from input training data set;
- Encoding the original point of data with the indices of the nearest code vectors in the codebook;
- Use index representation so as to reconstruct the data 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.

LBG Design Algorithm

1. Given $\mathcal{T}$. Fixed $\epsilon > 0$ to be a ``small" number.
2. Let $N = 1$ and

   $$\mathbf{c}_1^* = \frac{1}{M}\sum_{m=1}^{M}\mathbf{x}_m.$$

   Calculate

   $$D_{ave}^* = \frac{1}{Mk}\sum_{m=1}^{M}||\mathbf{x}_m - \mathbf{c}_1^*||^2.$$

3. **Splitting:** For $i = 1, 2, \ldots, N$, set

   $$\mathbf{c}_i^{(0)} = (1+\epsilon)\mathbf{c}_i^*,$$
   $$\mathbf{c}_{N+i}^{(0)} = (1-\epsilon)\mathbf{c}_i^*.$$

   Set $N = 2N$.
4. **Iteration:** Let $D_{ave}^{(0)} = D_{ave}^*$. Set the iteration index $i = 0$.
   i. For $m = 1, 2, \ldots, M$, find the minimum value of $||\mathbf{x}_m - \mathbf{c}_n^{(i)}||^2$, over all $n = 1, 2, \ldots, N$. Let n^* be the index which achieves the minimum. Set

   $$Q(\mathbf{x}_m) = \mathbf{c}_{n^*}^{(i)}.$$

   ii. For $n = 1, 2, \ldots, N$, update the codevector

   $$\mathbf{c}_n^{(i+1)} = \frac{\sum_{Q(\mathbf{x}_m)=\mathbf{c}_n^{(i)}}\mathbf{x}_m}{\sum_{Q(\mathbf{x}_m)=\mathbf{c}_n^{(i)}}1}$$

   iii. Set $i = i + 1$.
   iv. Calculate

$$D_{ave}^{(i)} = \frac{1}{Mk} \sum_{m=1}^{M} ||\mathbf{x}_m - Q(\mathbf{x}_m)||^2.$$

v. If $(D_{ave}^{(i-1)} - D_{ave}^{(i)})/D_{ave}^{(i-1)} > \epsilon$, go back to Step (i).

vi. Set $D_{ave}^{*} = D_{ave}^{(i)}$. For $n = 1, 2, \ldots, N$, set $\mathbf{c}_n^{*} = \mathbf{c}_n^{(i)}$ as the final codevectors.

5. Repeat Steps 3 and 4 until the desired number of codevectors is obtained. Once the codebook is generated, one can perform transformation using quantization

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.

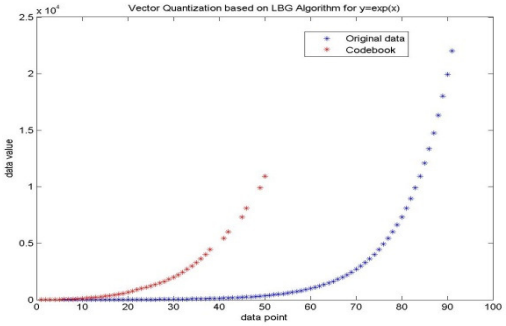

Figure 1: VQ based on LBG design Algorithm y=exp(x);

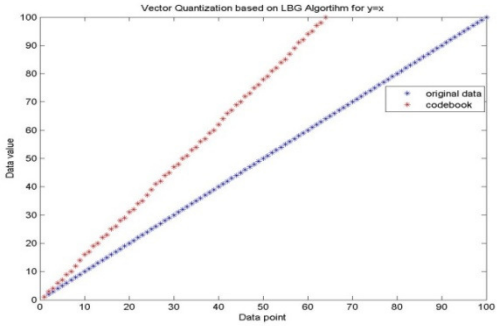

Figure 2: VQ based on LBG design Algorithm y=x;

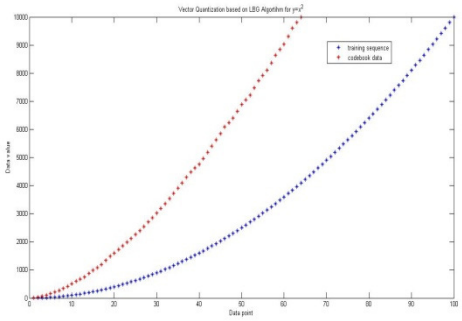

Figure 3: VQ based on LBG design Algorithm y=x2; y=sin(x);

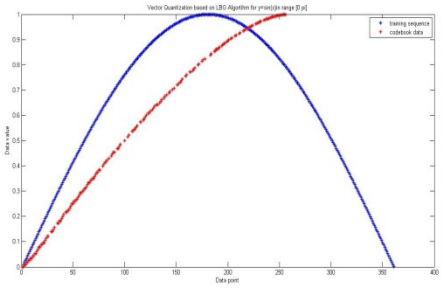

Figure 4: VQ based on LBG design Algorithm

Backup And recovery of Information Loss

Generally Privacy preserving data mining , we apply some techniques for modifying data and that modified data will be given to data miners.

In this paper we also concentrates keeping of original data as it is, so whenever data miners or owners of that data requires original data , they will get it by maintaining a backup copy of that data.

1. Taking all the data in to a DB1
2. We have to copy all the data in DB1 to DB2 (DB1=DB2)
3. Perform Data modification using LBG Design Algorithm and store the data in to DB1
4. Write the program to delete the all the contents in the backup of DB1,(so no except the users cannot access the information in DB1)
5. To see the actual details and to modify the details the admin should access the information in DB2
6. So, far the users know only DB1, data base owners can access to DB2

Bit Error Rate

In Data transmission, the number of **bit errors** is the number of received bits of a data stream over a communication channel that have been altered due to noise, interference, distortion or bit synchronization errors.

In our problem, we are transforming the original data to some other form using vector quantization. Hence we need to calculate the bit error rate for compressed data.
Always we try to minimise the bit error rate for accuracy

For example

Original data is
1 0 1 0 1 1 0 0 1 0

And after the transformation, the received data is
1 *1* 1 0 1 1 0 *1* 1 0
(Two errors are there, i.e, we are not receiving exact data only 80% accuracy is achieved because of two bit errors)

4. 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.

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