VECTOR QUANTIZATION FOR PRIVACY PRESERVING CLUSTERING IN DATA MINING

D.Aruna Kumari[1] , Dr.Rajasekhara Rao[2] and M.Suman[3]
[1,3] Department of Electronics and Computer Engineering, K.L.University, Vaddeswaram,Guntur
[1] aruna_D@kluniversity.in and [3] suman.maloji@gmail.com
[2] Department of Computer Science and Engineering, K.L.University, Vaddeswaram,Guntur
[2]rajasekhar.kurra@klce.ac.in

ABSTRACT

Large Volumes of personal data is regularly collected from different sources and analyzed by different types of applications using data mining algorithms , sharing of these data is useful 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 for preserving privacy. Quantization will be performed on training data samples it will produce transformed data set. This transformed data set does not reveal the sensitive data. And one can apply data mining algorithms on transformed data and can get accurate results by preserving privacy

KEYWORDS

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

1. INTRODUCTION

Privacy preserving data mining (PPDM) refers to the area of data mining that seeks to provide security for sensitive information from unsolicited or unsanctioned disclosure. Generally data mining techniques analyze the data; this analysis may be useful for decision making. On one hand it is useful for business people for taking correct decisions. On the other hand there is a Treat to privacy. now a day's privacy is becoming real since data mining techniques are able to predict high sensitive knowledge from data[3][1][2].

CDC is the best example for ppdm (Center for Disease Control and Prevention), there intension is to predict health threats, and to do so they require data from a range of sources (insurance companies, hospitals and so on), each of whom may be reluctant to share data. Insurance and hospital data is so sensitive , it may have high sensitive data ; if that is data revealed there may be treat to the party. Our aim is to provide knowledge by protecting the high sensitive data for decision making.

The term "privacy preserving data mining" was introduced (Agrawal & Srikant, 2000) and (Lindell & Pinkas, 2000). Agrawal and Srikant (2000) [3][21] devised a randomization algorithm that allows a large number of users to contribute their private records for efficient centralized data mining while limiting the disclosure of their values; Lindell and Pinkas (2000) invented cryptographic protocol for decision tree construction Other areas that influence the development of PPDM include cryptography and secure multiparty computation, distributed networks.

2. RELATED WORK

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

2.1 Perturbation: 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 are operating. 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.

2.2 Suppression

Privacy can be preserved by suppressing all sensitive data before performing any sort of computation or applying any data mining algorithm. For a given data base one can identify the sensitive attributes and suppress the attributes in some particular records. For a partial suppression, an exact attribute value can be replaced with a less informative value by rounding (e.g. $23.45 to $20.00), top-coding (e.g. age above 70 is set to 70), generalization (e.g. address to zip code), by using intervals (e.g. age 23 to 20-25, name 'Johnson' to 'J-K') etc. Often the privacy guarantee trivially follows from the suppression policy. Suppression cannot be used if data mining requires full access to the sensitive values

2.3 Cryptography:

The cryptographic approach to PPDM assumes[Binkas and Lindle] that the data is stored at several private parties, who agree to disclose the result of a certain data mining computation performed jointly over their data. The parties engage in a cryptographic protocol, i.e. they exchange messages encrypted to make some operations efficient while others computationally intractable. In effect, they "blindly" run their data mining algorithm. Classical works in secure multiparty computation such as Yao (1986) and Goldreich et al. (1987) show that any function F(x1, x2, …, xn) computable in polynomial time is also securely computable in polynomial time by n parties, each holding one argument, under quite broad assumptions regarding how much the parties trust each other. However, this generic methodology can only be scaled to database-sized arguments with significant additional research effort. [21]

• blocking, which is the replacement of an existing attribute value with a "?",

• aggregation or merging which is the combination of several values into a coarser category,

• swapping that refers to interchanging values of individual records, and

3.PROPOSED APPROACH

3.1 Privacy preserving clustering :

The goal of privacy-preserving clustering is to protect the underlying attribute values of objects subjected to clustering analysis. In doing so, the Privacy of individuals would be protected. The problem of privacy preservation in clustering can be stated as follows as in [6][7]: Let *D* be a relational database and *C* a set of clusters generated from *D*. The goal is to transform *D* into *D'* so that the following restrictions hold:

1. A transformation T when applied to *D* must preserve the privacy of individual records, so that the released database *D'* conceals the values of confidential attributes, such as salary, disease

diagnosis, credit rating, and others.

2. The similarity between objects in *D'* must be the same as that one in *D*, or can have little bit difference
Between them which was occurred by the transformation process.

Transformed database *D'* looks very different from *D*, the clusters in *D* and *D'* should be as close as possible since the distances between objects are preserved or marginally changed.

That transformation can done by Vector quantization

Our work is based on piecewise Vector Quantization method and is used as non dimension reduction method. It is modified form of piecewise vector quantization approximation which is used as dimension reduction technique for efficient time series analysis in [7].

3.2 Vector Quantization:

Vector Quantization (VQ) is an efficient technique for data compression and it is used image processing, signal processing,…etc. it is efficient compared to scalar quantization[14]. In other words, the objective of VQ is the representation of vectors $X \subseteq Rk$ by a set of reference vectors $CB = \{C1;C2; : : : ;CN\}$ in Rk in which Rk is the k-dimension Euclidean space. CB is a codebook which has a set of reproduction codewords and $Cj = \{c1; c2; : : : ; ck\}$ is the j-th codeword. The total number of codewords in CB is N and the number of dimensions of each codeword is k.

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;
- reconstructing or decoding the signal with the index of the codebook.

Since we have not to reconstruct the original data so above two steps only are involved such that it is difficult to reconstruct the original data thus preserving privacy but it should be represented by most approximate data such that similarity between data is preserved which can lead to accurate clustering result. Moreover rather than indices we use direct code vector for encoding. Huge training dataset will be taken. The codebook is generated from these training vectors by the clustering techniques.

In the encoding procedure, an original data is divided into several k-dimension vectors and each vector is encoded by the index of codeword During the decoding procedure, the receiver uses the same codebook to translate the index back to its corresponding codeword for reconstructing the original one. This is what exactly happening in Vector quantization[13][15][16].

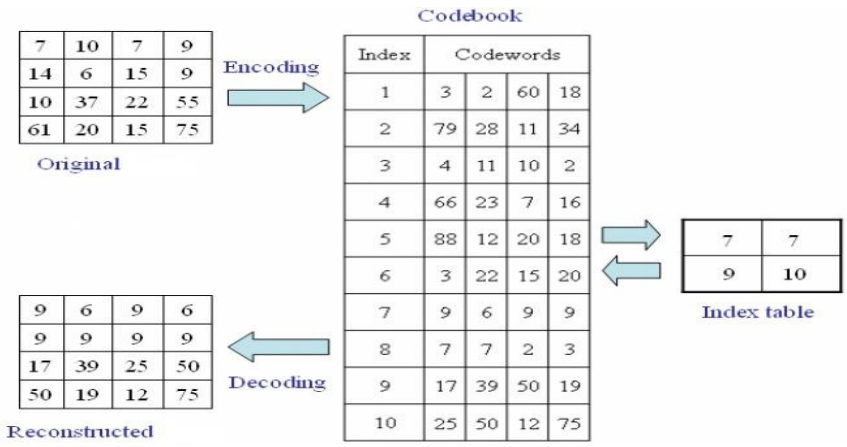


FIGURE 1. An Example of Encoding and Decoding by VQ

As per as Our PPDM is considered , we do not required to get back the original one, so we can follow the first two steps i.e, Constructing codebook and encoding .Hence privacy is preserved.

3.3 Code Book Generation using K-means Clustering

Step1: The training vectors are grouped into M clusters based on the distance between the code vectors and the training vectors using distance measure

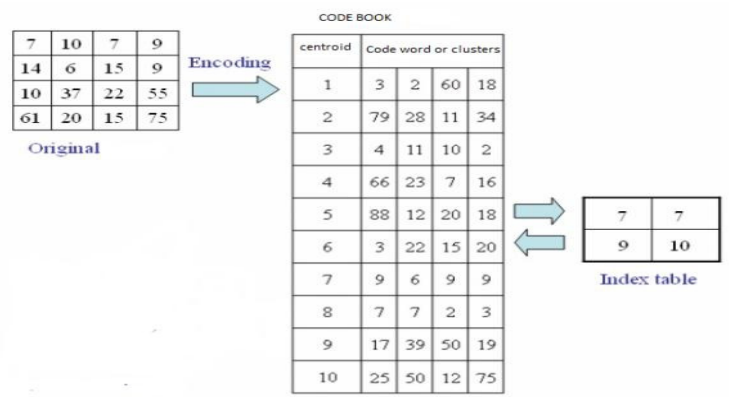
Step2: Compute the sum vector for every cluster by adding the corresponding components of all the training vectors that belong to the same cluster using the equation (16).

Step3: Determine the centroid for each cluster. And calculate difference between old centroid and new centroid, if it is greater than 1 perform step 4.

Step4: Replace the existing codevector with the new centroid to form the revised codebook.

Step5: Repeat the above steps till the complete codebook is generated.

Privacy Preserving Using Vector Quantization : First we construct Code book from Huge training samples and then using encoding we will get transformed data set hence privacy will be preserved. For achieving privacy preserving one can follow first two stages. So that privacy is preserved.



7	10	7	9
14	6	15	9
10	37	22	55
61	20	15	75

centroid	Code word or clusters			
1	3	2	60	18
2	79	28	11	34
3	4	11	10	2
4	66	23	7	16
5	88	12	20	18
6	3	22	15	20
7	9	6	9	9
8	7	7	2	3
9	17	39	50	19
10	25	50	12	75

7	7
9	10

FIGURE 2: An Example for encoding using VQ , Preserves Privacy

In the figure 2. We have performed only up to encoding; there by one cannot predict exact data other than the centroids or cluster centers. One of the important point in VQ is the construction of Code book , but code book generation is a time consuming process. reducing computation time for VQ is an important issue.

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.

5. 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]clustering in data mining".

[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, and C. S. Yang, A Fast VQ Codebook Generation Algorithm via Pattern Reduction, *Pattern Recognition Letters*, vol. 30, pp. 653{660, 2009}

[14]C. W. Tsai, C. Y. Lee, M. C. Chiang Kurt Thearling, Information about data mining and analytic technologies http://www.thearling.com/

[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]Quantization: A Review", IEEE Transactions on Communications, Vol. 36, No. 8, August 1988.

[19]Berger T, "Rate Distortion Theory", Englewood Cliffs, Prentice-Hall,NJ, 1971.

[20] A.Gersho and V.Cuperman, "Vector Quantization: APattern Matching Technique for Speech Coding", IEEE Communications, Mag., pp 15-21, 1983.

[21]"Privacy Preserving Data Mining - IBM Research: Almaden: San Jose

[22]D.Aruna Kumari, Dr.K.Rajasekhara rao, M.suman "Privacy Preserving Clustering in DDM using Cryptography"in TJ-RJCSE-IJ-06

Authors

D.Aruna Kumari Assoc.professor ECM Dept,K.L.University has 7 years of experience in teaching working in the area of Data Mining and has published around 30 papers in various conferences/journals. Life member CSI

Dr.K.Rajasekhara Rao Professor ECM Dept, K.L. University has 25 years experience in teaching/management. Research area is Software engineering, Data Mining & Embedded Systems and has published around 45 papers in various conferences/journals.CSI life member and chairman for AP student committee

M.Suman Assoc.professor ECM Dept, and Assistant registratr K.L.University working in the area of Speech Processing and has published around 30 papers in various conferences/journals.CSI life