PRIVACY PRESERVING CLUSTERING IN DATA MINING USING VQ CODE BOOK GENERATION

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

Huge 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 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 original data. Hence privacy is preserved.

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. Most traditional data mining techniques analyze and model the dataset statistically, in aggregation, while privacy preservation is primarily concerned with protecting against disclosure of individual data records. Treat to privacy is becoming real since data mining techniques are able to predict high sensitive knowledge from unclassified data[1][2].

Common examples arise in health science, where data may be held by multiple parties: commercial organizations (such as drug companies, or hospitals), government bodies (such as the Food and Drug Administration) and non-government organizations (such as charities)[1]. Each organization is bound by regulatory restrictions (for instance privacy legislation), and corporate

Sundarapandian et al. (Eds): CoNeCo,WiMo, NLP, CRYPSIS, ICAIT, ICDIP, ITCSE, CS & IT 07,
pp. 317–323, 2012. © CS & IT-CSCP 2012 DOI : 10.5121/csit.2012.2430

requirements (for instance on distributing proprietary information that may provide commercial advantage to competitors). In such a case, an independent researcher may not receive access to data at all, while even members of one of these organizations sees an incomplete view of the data. However, data from multiple sources may be needed to answer some important questions. A classical example occurs for an organization like the CDC (Center for Disease Control and Prevention), who are mandated with detecting potential 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.

The term "privacy preserving data mining" was introduced in papers (Agrawal & Srikant, 2000) and (Lindell & Pinkas, 2000). These papers considered two fundamental problems of PPDM, privacy preserving data collection and mining a dataset partitioned across several private enterprises. Agrawal and Srikant (2000) 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 a cryptographic protocol for decision tree construction over a dataset horizontally partitioned between two parties. These methods were subsequently refined and extended by many researchers worldwide.

Other areas that influence the development of PPDM include cryptography and secure multiparty computation (Goldreich, 2004) (Stinson, 2006), database query auditing for disclosure detection and prevention (Kleinberg et al. 2000) (Dinur & Nissim, 2003) (Kenthapadi et al. 2005), database privacy and policy enforcement (Agrawal et al. 2002) (Aggarwal et al. 2004), database security (Castano et al. 1995), and of course, specific application domains.

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 Supression

Privacy can be preserved by suppressing all sensitive data before any disclosure or computation occurs. For a given data based one can 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. However, the analysis may be difficult if the choice of alternative suppressions depends on the data being suppressed, or if there is dependency between disclosed and suppressed data. Suppression cannot be used if data mining requires full access to the sensitive values

2.3 Cryptography:

The cryptographic approach to PPDM assumes 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.

• 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
• sampling, which refers to releasing data for only a sample of a population.

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 just slightly altered by the transformation process. Although the 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 and simple approach for data compression. Since it is

simple and easy to implement, VQ has been widely used in different applications, such as pattern recognition, image compression, speech recognition, face detection and so on [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;
- Using an index representation to reconstruct the signal by looking up in 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 samples will be taken. The representative codebook is generated from these training vectors by the clustering techniques. Then In the encoding procedure, an original data is divided into several k-dimension vectors and each vector is encoded by the index of codeword by a table look-up method. 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.

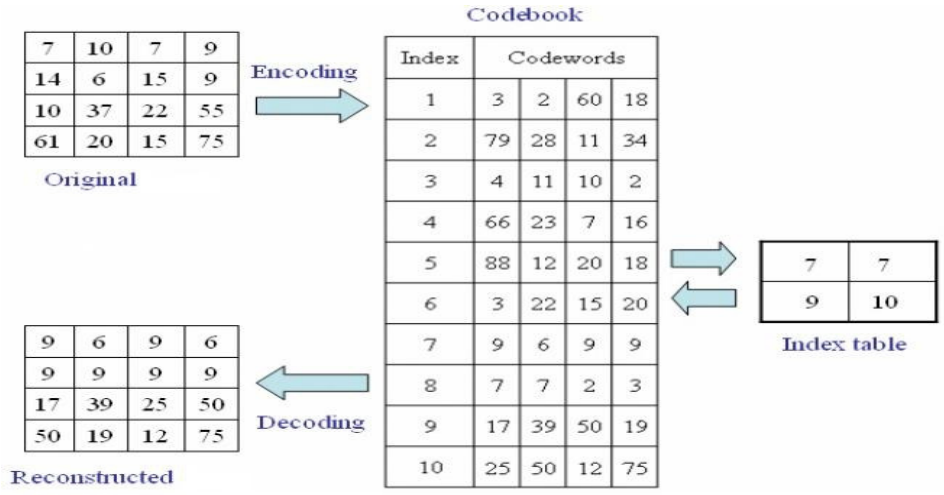


FIGURE 1. An Example of Encoding and Decoding by VQ

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.

3.3 Code Book Generation using K-means Clustering

Step1: The training vectors are grouped into M clusters based on the distance between the codevectors and the training vectors using the equation (15) and (16).

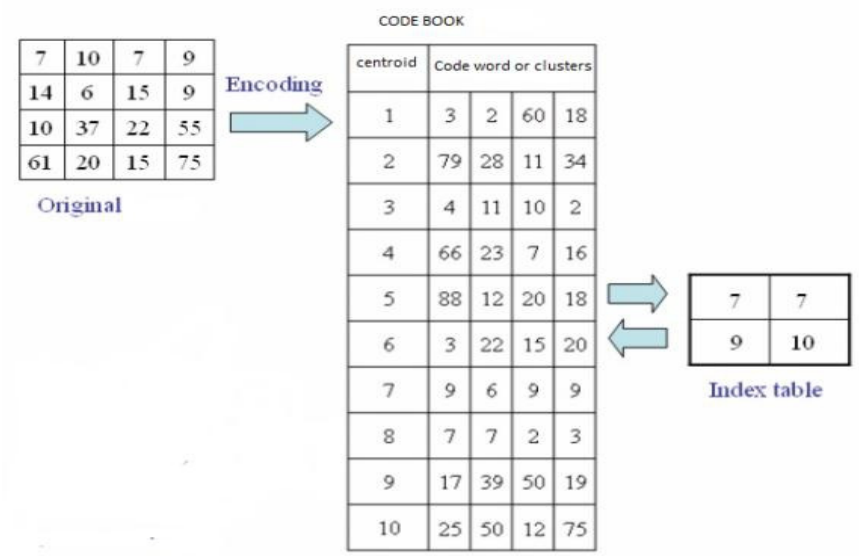
Step2: Compute the sum vector for every cluster by adding the corresponding components of all the training vectors that belong to the same cluster using the equation (16).
Step3: Compute the centroid for each cluster by dividing the individual components of the sum vector by the cluster strength ni using the equation (125.

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

Step5: Repeat the steps 1 through 4 till the codebooks of the consecutive iterations converge.

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.

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.

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 member