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 10, October -2016

PRIVACY PRESERVING DM USING K-MEANSLBG FOR VECTOR QUNTIZATION (KVQ) AND NOISE ADDITION

[1]Brijal Patel, [2] Dimple Kanani

[1,2]Assistant professor, IT department, Vadodara Institute of Technology, Vadodara, India

Abstract: *Perturbation method is a very important technique in privacy preserving data mining. In this technique, the tradeoff is in between loss of information and preservation of privacy. The question is, how much are the users willing to compromise their privacy? This is a choice that changes from individual to individual. The data comes from a heterogeneous environments including financial, library,shopping, medicaland telephonerecords.As it is possible due to the rapid growth in database,computing, andnetworking technologiesso such data can be integrated and analyzed digitally. In order to share data while preserving privacy data owner must come up with a solution which achieves the dual goal of privacy preservation as well as accurate clustering result. Trying to give solution for this we implemented vector quantization approach piecewise on the datasets which segmentize each row of datasets and quantization approach is performed on each segment using K means clustering algorithm.*

Keywords: *Privacy, Data, PVQ, Preserve.*

I. INTRODUCTION

Data mining is a technique that deals with the extraction of hidden predictive information from large database. It uses sophisticated algorithms for the process of sorting through large amounts of data sets and picking out relevant information.Data mining evolvedlong process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in pattern access, and more recently, generated tools and technologies that allow to move their data in real time.

A. Data Mining Functions

- Classification: Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify loan applicants as low, medium, or high credit risks.The database contains one or more attributes that denote the class of a tuple and these are known as predicted attributes whereas the remaining attributes are called predicting attributes.

- Associations:Association is a data mining function that discovers the probability of the co-occurrence of items in a collection. The relationships between co-occurring items are expressed as association rules. Association rules are often used to analyze sales transactions. Associations can involve any number of items on either side of the rule.Association rule mining is a procedure which is meant to find frequent patterns, associations, correlations, or causal structures from data sets found in heterogeneous databases such as transactional databases, relational databases and other forms of data repositories.

- Clustering/Segmentation:Clustering is a process of partitioning a set of data (or objects) into a set of meaningful sub-classes, called clusters. Help users understand the natural grouping or structure in a data set. Clustering according to similarity is a very powerful technique, the key to it being to translate some intuitive measure of similarity into a quantitative measure. When learning is unsupervised then the system has to discover its own classes i.e. the system clusters the data in the database. The system has to discover subsets of related objects in the training set and it has to find descriptions of each subsets.

II.PRIVACY PRESERVING DATA MINING

Most of the techniques for PPDM uses modified version of standard data mining algorithms, where the modifications usually using well known cryptographic techniques ensure the required privacy for the application for which the technique was designed. In most cases, the constraints for PPDM are preserving accuracy of the data and the generated models and the performance of the mining process while maintaining the privacy constraints. The several approaches used by PPDM can be summarized as below:

1. The data is altered before delivering it to the data miner.
2. The data is distributed between two or more sites, which cooperate using a semi-honest protocol to learn global data mining results without revealing any information about the data at their individual sites.
3. While using a model to classify data, the classification results are only revealed to the designated party, who does not learn anything else other that the classification results, but can check for presence of certain rules without revealing the rules.

A number of techniques such as randomization and *k*-anonymity have been suggested in recent years in order to perform privacy-preserving data mining. Furthermore, the problem has been discussed in multiple communities such as the database community, the disclosure control community and the cryptography community. In some cases, the different communities have explored parallel lines of work which are quite similar.

Problem Definition:

The problem of privacy preservation in clustering can be stated as follows as in: 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:

- 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.

- 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.

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.

III. PROPOSED METHODOLOGY

NOISE ADDITION	VECTOR QUNTIZATION
1.Extract sensitive attributes fron given dataset.	1. Input dataset with sensitive information.
2.Apply min-max normalization formula on sensitive attributes for noise addition as $v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new$	2. Dataset is segmented as : 1st Segment Y1 = x1 x2 x3……xL 2nd Segment Y2 = xL+1 xL+2 xL+3……x2L 3rd Segment Y3 = x2L+1 x2L+2 x2L+3……x3L With Segment Yw = x(w-1)L+1 x(w-1)L+2 x(w-1)L+3 …xwL
3. Calculate sensitive attribute's value as: $f(x) = \sum_{n=1}^{N} D(n)\left(\cos\frac{\pi(2n-1)(k-1)}{2N}\right)$ Where k = 1,2……N $D(n) = [S]_{nx3}$	3. Generate codebook using k-means algorithm. use the distance measures to calculate similarity and dissimilarity.

4. Multiply f(x) with $\frac{1}{\sqrt{n}}$ for k=1 or multiply with $\sqrt{\frac{2}{N}}$ for 2	4. Quantization using data transformation: 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.
5. Crate perturbed dataset D' by replacing sensitive attribute in original dataset D with f(x).	5. Dataset reformation: 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.
6. Apply k-Mean clustering algorithm with different values of k on original dataset D having sensitive attribute S.	6 Comparison for accuracy from distortion in data Clustering by K means is performed on original dataset and result received (R1) 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.
7 Apply k-Mean clustering algorithm with different values of k on perturbed dataset D' having perturbed sensitive attribute P.	
8 Create cluster membership matrix of results from step 5 and step 6 and analyze.	

IV. RESULTS & DISCUSSIONS

We have implemented above proposed algorithm using 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 centroid.

In this experiment, we compare the accuracy of privacy preserving kNN classification against a dataset. In these experiments, we run the nearest neighbor selection step for one round, with the initial probability P0 = 1 and the randomization factor d = 0:5.In figure 4.2 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 centroid.

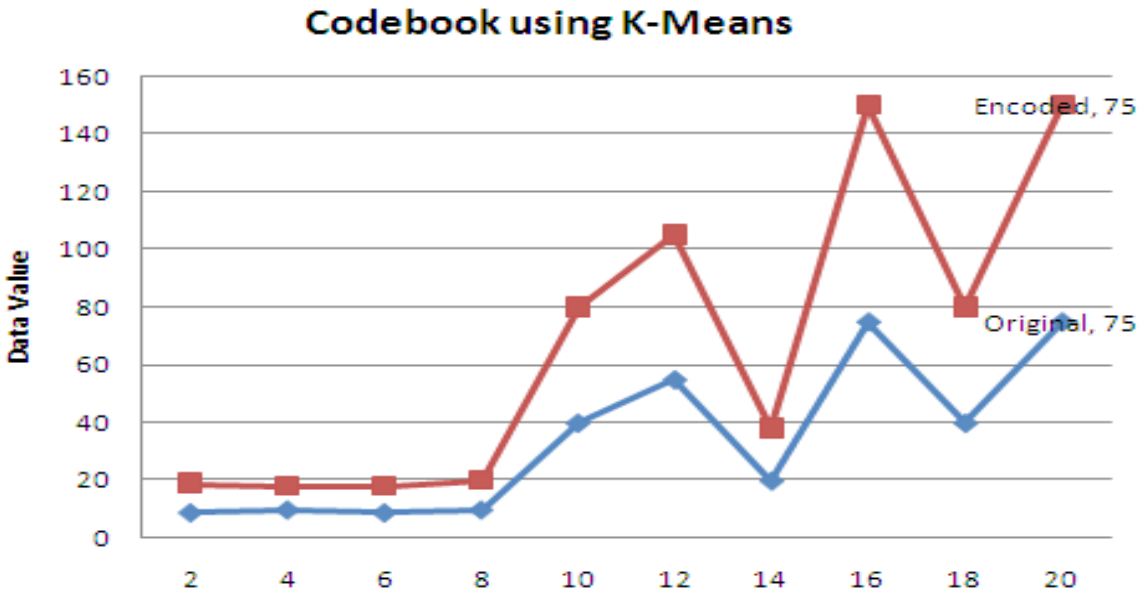

Figure 4.1 Representation of original data set and quantized data

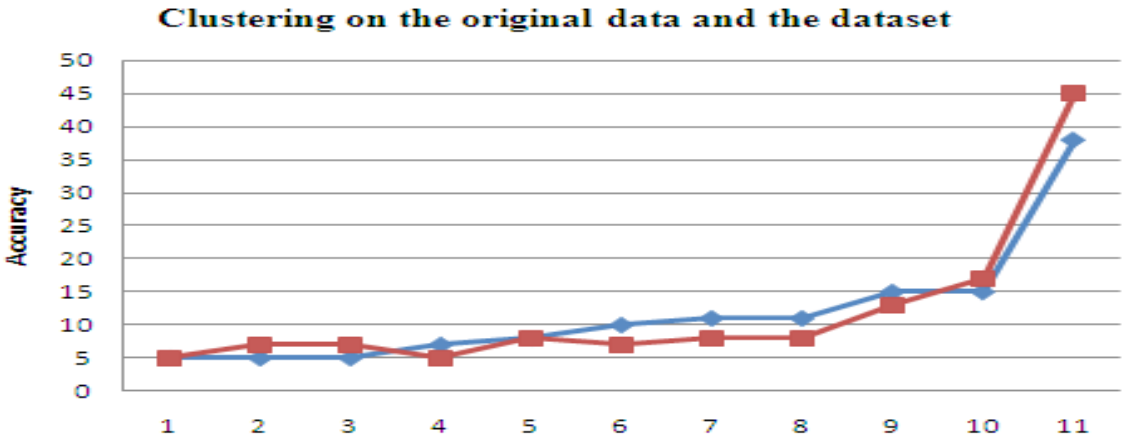

Figure 4.2 Clustering on original data and quantized dataset for accuracy using K-Means

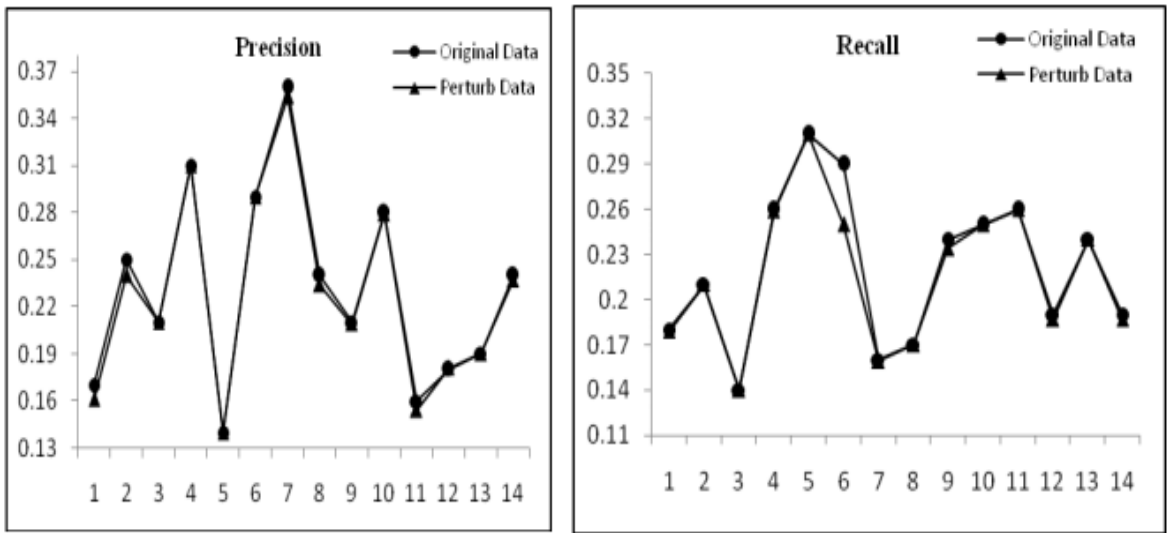

Figure 4.3 Accuracy on attribute Time_Duration in Basketballdataset (w=2000 and k=5)

The Vector Quantization techniques are efficientlyused to increase the performance of the speech recognition system. The recognition accuracy obtained using K-meansLBG for vector quantization algorithm is better as compared to K-means and LBG algorithm. The average recognition accuracy of K-meansLBG for vector quantization algorithm is more than 2.55% using K-means algorithm while the average recognition accuracy of K-meansLBG for vector quantization algorithm is more than 1.41% using LBG algorithm.

For noise addition we had taken multiple window size for different number of sensitive attributes. From that we noticed that if window size is 2000 and no. of sensitive attributes are 3 than there is highest accuracy of data that is 99.41%

V. CONCLUSIONS

This work is based on vector quantization and noise addition. Finally we would like conclude that Efficiency depends on the code book generation. So we get efficiency more than 2.55% if we use vector quantization. In this work, we have considered, for the first time, the issue of providing efficiency in privacy preserving mining. A new K-MeansLBG algorithm that is specifically designed to minimize this side-effect through the application of symbol specific distortion. And if we use noise addition for accuracy, we get maximum accuracy using precision and recall which removes outliers.

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. Flavius L. Gorgônio and José Alfredo F. Costa"Privacy- Preserving Clustering on Distributed Databases:A Review and Some Contributions
10. 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.
11. 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}
12. C. W. Tsai, C. Y. Lee, M. C. Chiang Kurt Thearling, Information about data mining and analytic technologies http://www.thearling.com/
13. 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.
14. 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
15. Vassilios S. Verykios, Elisa Bertino, Igor Nai Fovino Stateof- the-art in Privacy Preserving Data Mining in SIGMOD Record, Vol. 33, No. 1, March 2004.
16. Brijal Patel H ,Ankur N shah, "Privacy Preserving in DM using min-max normalization and Noise addition", International journal of advanced Engineering and research Development, Vol. 2, Issue 10, October-2015.

