VECTOR QUANTIZATION FOR PRIVACY PRESERVING DATA MINING

1D.ARUNA KUMARI, 2K.RAJASEKHARA RAO, 3M.SUMAN, 4HIMA BINDU Y, 5SRIVIDYA B, 6KAMALA G

1,3 Department of Electronics and Computer Engineering, Associate professors ,CSI Life Member K.L.University, Vaddeswaram, Guntur
2 Professor and Dean(Administration), Department of CSE, CSI life member
3,5,6 IV/IV B.Tech ECM K.L.University, Vaddeswaram, Guntur

Abstract: Privacy preserving data mining is becoming increasingly important issue as it predicts high sensitive information. In this paper we provide new dimensions of privacy preserving data mining i.e, transformation using vector quantization. We will show reconstruction based technique for numerical data. Finally the performance of the proposed techniques is evaluated using accuracy and distortion parameters.

Keywords: Privacy preserving, reconstruction, quantization.

I. INTRODUCTION

Huge volumes of detailed personal data are regularly collected and analyzed by applications. Such Data include shopping habits, criminal records, medical history, credit records, among others. Analyzing such data opens new threats to privacy .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. 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].

II. CLASSIFICATION OF PRIVACY PRESERVING DATA MINING

There are many approaches for privacy preserving data mining. Privacy preserving data mining techniques can be classified based on the following dimensions[11].

1. Data distribution
2. Data modification
3. Data mining algorithm
4. Data or rule hiding
5. Privacy preservation

First dimension refers to data distribution; data can be distributed vertically or horizontally over the systems. Second dimension refers to modifying the original data to other form, so that we can prevent de-identification of sensitive data. There are several methods are there for data modification like randomization, swapping, sampling, anonymity, blocking,…etc

Third dimension is Data mining algorithm, when mining is performed on data we could be able to preserve privacy of individuals. Fourth dimension refers to Hiding, part of data or result of data mining … can be hided. Fifth dimension is the most important issue i.e, providing privacy during data mining .This paper mainly concentrates on fifth dimension

III. PROPOSED APPROACH

Vector quantization works based on rounding off.

1. Original Data
2. Constructing codebook
3. Encoding original data with code book
4. Decoding

Vector quantization (VQ) is generally used for data compression. In previous days, the design methodology of a vector quantizer (VQ) is treated as a big problem in terms of the need for multi-dimensional integration. Linde, Buzo, and Gray (LBG) Introduced an algorithm for Vector quantization design based on training sequence. A VQ that is designed based on this algorithm are referred as LBG-VQ.

Figure 3.1 : Block diagram of Vector Quantizer
Algorithm 1.a

Heuristic-based techniques, algorithm 1.a encodes only selected values that minimize the utility loss rather than all available values.

Algorithm 1.b

Algorithm 1.b is reconstruction-based techniques where the original distribution of the data is reconstructed from codebook.
CONCLUSION AND FUTURE SCOPE

This work gives a different approach of using vector quantization for privacy preserving data mining. This work shows analytically and experimentally that Privacy-Preserving data mining is to some extent possible using vector quantization approach. To support this work, water treatment dataset available on UCI Machine Learning Repository was taken and performed experiments on it. Performance is also evaluated by taking into account two important parameter: distortion and Fmeasure (quality of data mining results).

As future work new and effective quantization method can be used rather than LBG approach that we have used. K nearest neighbor approach is one of the approach which can give better result.

REFERENCES

2. Alexandre Evfimievski, Tyrone Grandison Privacy Preserving Data Mining, IBM Almaden Research Center 650 Harry Road, San Jose, California 95120, USA.
7. Binit Kumar Sinha “Privacy preserving clustering in data mining”.