ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

Privacy Preserving Data Mining Using Piecewise Vector Quantization (PVQ)

[I]S. Sasikala, [II]S. Nathira Banu
[I]HOD, Dept. of UG Computer Science, [II]Research Scholar
[I,II]Saraswathi Thyagaraja College, Pollachi.

Abstract
Over the last twenty years, there has been an extensive growth in the amount of private data collected about individuals. This data comes from a number of sources including medical, financial, library, telephone, and shopping records. Such data can be integrated and analyzed digitally as it's possible due to the rapid growth in database, networking, and computing technologies. On the one hand, this has led to the development of data mining tools that aim to infer useful trends from this data. But, on the other hand, easy access to personal data poses a threat to individual privacy. Here, we provide the piecewise quantization approach for dealing with privacy preserving clustering.
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 which later are again united to form a transformed data set. Some experimental results are presented which tries to finds the optimum value of segment size and quantization parameter which gives optimum in the tradeoff between clustering utility and data privacy in the input dataset.

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 tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. With the amount of data doubling each year, more data is gathered and data mining is becoming an increasingly important tool to transform this data into information. Long process of research and product development evolved data mining. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery.

A. Data Mining Functions

Data mining methods may be classified by the function they perform or according to the class of application they can be used in. Some of the main techniques used in data mining are described in this section.

1. Classification

Data mine tools have to infer a model from the database, and in the case of supervised learning this requires the user to define one or more classes. 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. A combination of values for the predicted attributes defines a class.
When learning classification rules the system has to find the rules that predict the class from the predicting attributes so firstly the user has to define conditions for each class, the data mine system then constructs descriptions for the classes. Basically the system should given a case or tuple with certain known attribute values be able to predict what class this case belongs to. Once classes are defined the system should infer rules that govern the classification therefore the system should be able to find the description of each class.

2. Associations

Given a collection of items and a set of records, each of which contain some number of items from the given collection, an association function is an operation against this set of records which return affinities or patterns that exist among the collection of items. These patterns can be expressed by rules such as "72% of all the records that contain items A, B and C also contain items D and E." The specific percentage of occurrences (in this case 72) is called the confidence factor of the rule. Also, in this rule, A,B and C are said to be on an opposite side of the rule to D and E. Associations can involve any number of items on either side of the rule.

3. Sequential/Temporal patterns

Sequential/temporal pattern functions analyse a collection of records over a period of time for example to identify trends. Where the identity of a customer who made a purchase is known an analysis can be made of the collection of related records of the same structure (i.e. consisting of a number of items drawn from a given collection of items). The records are related by the identity of the customer who did the repeated purchases. Such a situation is typical of a direct mail application where for example a catalogue merchant has the information, for each customer, of the sets of products that the customer buys in every purchase order. A sequential pattern function will analyse such collections of related records and will detect frequently occurring patterns of products bought over time. A sequential pattern operator could also be used to discover for example the set of purchases that frequently precedes the purchase of a microwave oven.

4. Clustering/Segmentation

Clustering and segmentation are the processes of creating a partition so that all the members of each set of the partition are

similar according to some metric. A cluster is a set of objects grouped together because of their similarity or proximity. Objects are often decomposed into an exhaustive and/or mutually exclusive set of clusters.

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 then it has to find descriptions that describe each of these subsets.

II. Privacy Preserving Data Mining

Privacy preserving data mining (PPDM) has emerged to address this issue. 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.

The problem of privacy-preserving data mining has become more important in recent years because of the increasing ability to store personal data about users, and the increasing sophistication of data mining algorithms to leverage this information. 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 statistical disclosure control community and the cryptography community. In some cases, the different communities have explored parallel lines of work which are quite similar.

The key directions in the field of privacy-preserving data mining are as follows:

i) Privacy-Preserving Data Publishing:

These techniques tend to study different transformation methods associated with privacy. These techniques include methods such as randomization, *k*-anonymity, and *l*-diversity. Another related issue is how the perturbed data can be used in conjunction with classical data mining methods such as association rule mining. Other related problems include that of determining privacy-preserving methods to keep the underlying data useful (utility-based methods), or the problem of studying the different definitions of privacy, and how they compare in terms of effectiveness in different scenarios.

ii) Changing the results of Data Mining Applications to preserve privacy

In many cases, the results of data mining applications such as association rule or classification rule mining can compromise the privacy of the data. This has spawned a field of privacy in which the results of data mining algorithms such as association rule mining are modified in order to preserve the privacy of the data. A classic example of such techniques are association rule hiding methods, in which some of the association rules are suppressed in order to preserve privacy.

iii) Query Auditing:

Such methods are akin to the previous case of modifying the results of data mining algorithms. Here, we are either modifying or restricting the results of queries.

iv) Cryptographic Methods for Distributed Privacy:

In many cases, the data may be distributed across multiple sites, and the owners of the data across these different sites may wish to compute a common function. In such cases, a variety of cryptographic protocols may be used in order to communicate among the different sites, so that secure function computation is possible without revealing sensitive information.

v) Theoretical Challenges in High Dimensionality:

Real data sets are usually extremely high dimensional and this makes the process of privacy-preservation extremely difficult both from a computational and effectiveness point of view. In, it has been shown that optimal *k*-anonymization is NP-hard. Furthermore, the technique is not even effective with increasing dimensionality, since the data can typically be combined with either public or background information to reveal the identity of the underlying record owners.

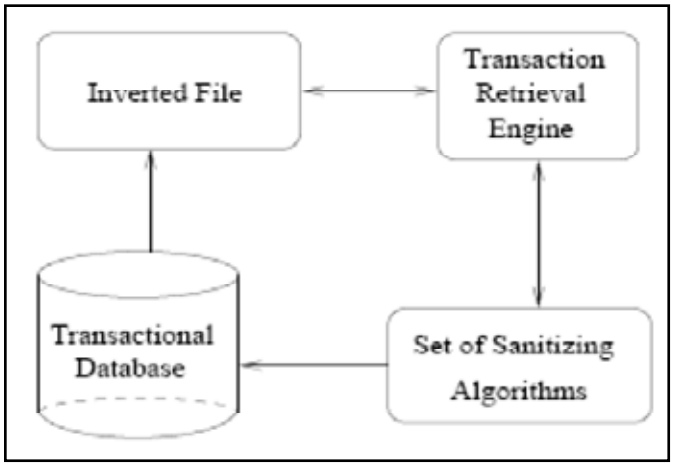

Fig .2.1: - Framework for Privacy Preserving Data Mining

However, unlike as is the case for statistical databases, another objective of collective privacy preservation is to preserve strategic patterns that are paramount for strategic decisions, rather than minimizing the distortion of all statistics (e.g., bias and precision). In other words, the goal here is not only to protect personally identifiable information but also some patterns and trends that are not supposed to be discovered.

A. Problem Definition

The goal of privacy-preserving clustering is to protect the underlying attribute values of objects subjected to clustering

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

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: 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

Vector Quantization is widely used in signal compression and coding. It is a lossy compression method based on principle block coding. We have used it in privacy preserving by approximating each point (row of data) to the other with the help of vector quantization approach (VQ).There is no data compression but there is quantization of data so that privacy is preserved.

Step 1: - Input

Input is dataset which is stored in file which contains sensitive information which is to be preserved such that there is less information loss and hence good clustering result.

Each point (row) of data is sequence of real value X = x1 x2 x3…….xn . Dataset contain "m" row of data.

Step 2: - Segmentation

Dataset is segmented into w datasets by decomposing each row of data into w segments each of length L. Let each row of data is represented by X = x1 x2 x3…….xn Then it is segmentize into w segments as given below.

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

Dataset D is decomposed into D1 D2 D3 ……DW dataset where each row of each dataset is of length L. If total number of attributes in original is not perfectly divisible by L then extra attributes is added with zero value which does not affect the result and later it is removed at step 5.

Step 3: - Clustering for Codebook Generation

In order to generate codebook which is helpful in data transformation, K means clustering algorithm is used. The algorithm accepts two inputs. The data itself (in step 1), and "k", the number of clusters. The aim of K-means (or clustering) is this:

We want to group the items into k clusters such that all items in same cluster are as similar to each other as possible. And items not in same cluster are as different as possible. We use the distance measures to calculate similarity and dissimilarity.

One of the important concepts in K-means is that of centroid. Each cluster has a centroid. One can consider it as the point that is most representative of the cluster. Equivalently, centroid is point that is the "center" of a cluster.

Step 4: - Data Transformation by Quantization

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. That is the point is replaced by the cluster centroid in which it falls.

Step 5: - Reformation of Dataset

Each segment Yi of row data X formed as segmentation step in step 2 is transformed into Zi by step 4. Now all the w 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.

Step 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.

A. Proposed Modified K-meansLBG Algorithm

The proposed algorithms objective is to overcome the limitations of LBG algorithm and K-means algorithm. The proposed modified K-MeansLBG algorithm is the combination of advantages of LBG algorithm and K-means algorithms. The K-MeansLBG algorithm is described as given below:

Step 1: Randomly select N training data vectors as the initial code vectors.

Step 2: Calculate the no. of centroid.

Step 3: Double the size of the codebook by splitting.

Step 4: Nearest-Neighbor Search.

Step 5: Find Average Distortion.

Step 6: Update the centroid till there is no change in the clustering centroids, terminate the program otherwise go to step 1.

IV. Results & Discussions

We have implemented above LBG algorithm using Matlab 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 centroid. Generally Privacy preserving data mining, we apply some techniques for modifying data and that modified data will be given to data miners. Here 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.

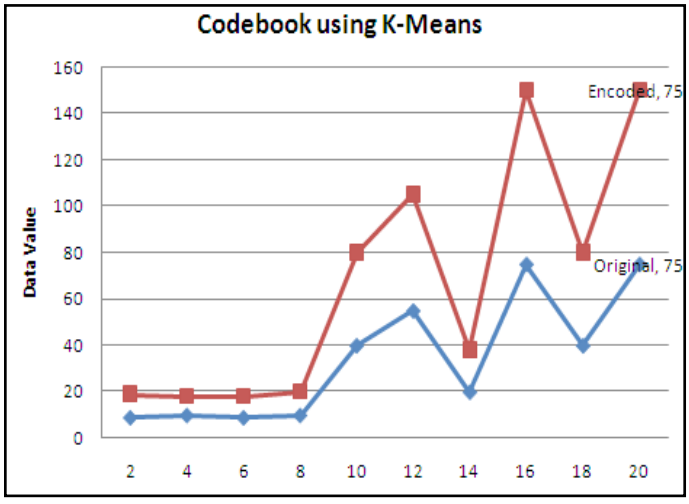

Fig. 4.1: - Representation of original data set and quantized data set

An important design goal for privacy preserving classification algorithms is their efficiency. We would ideally like the complexity of the privacy preserving classification algorithm to be of the same order as the complexity of a non-privacy preserving classification algorithm.

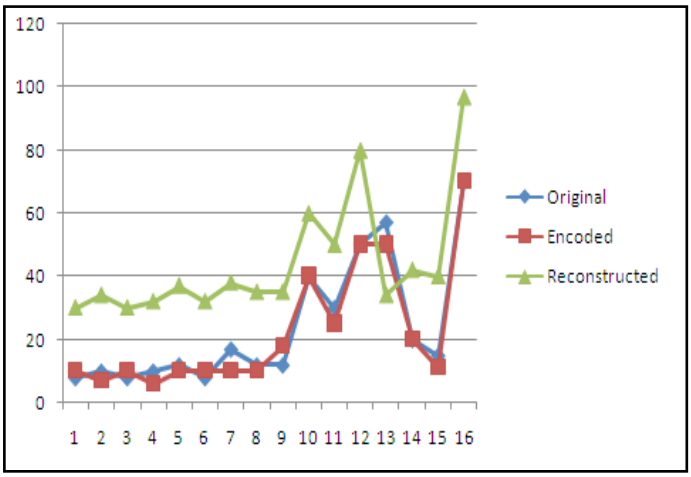

Fig. 4.2: After reconstructed the dataset

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.

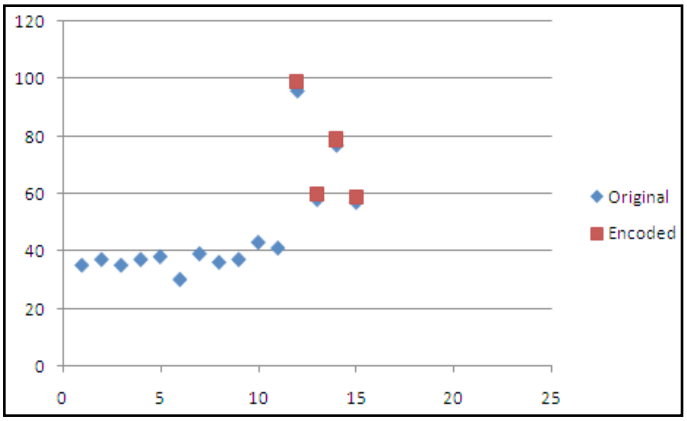

Fig. 4.3: Clustering on Encoded Data

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 centroid.

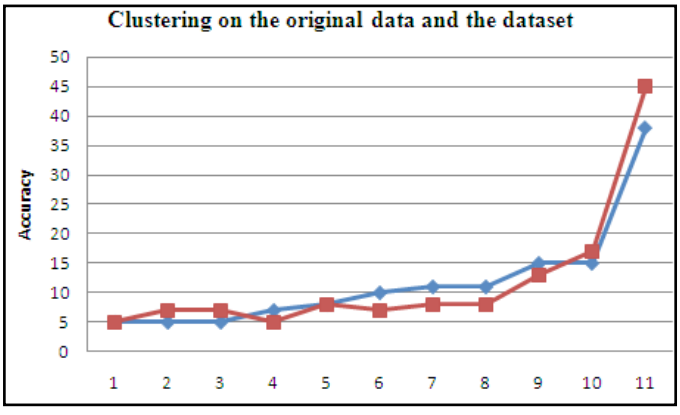

Fig 4.4: Clustering on original data and quantized dataset for accuracy using K-Means

The relative accuracy of our privacy preserving algorithm is determined by the accuracies of its two steps. The accuracy of the privacy preserving nearest neighbor selection step is determined by the accuracy of the PP - TopK algorithm. As PP - TopK is a randomized algorithm; we can only present probabilistic guarantees on its accuracy.

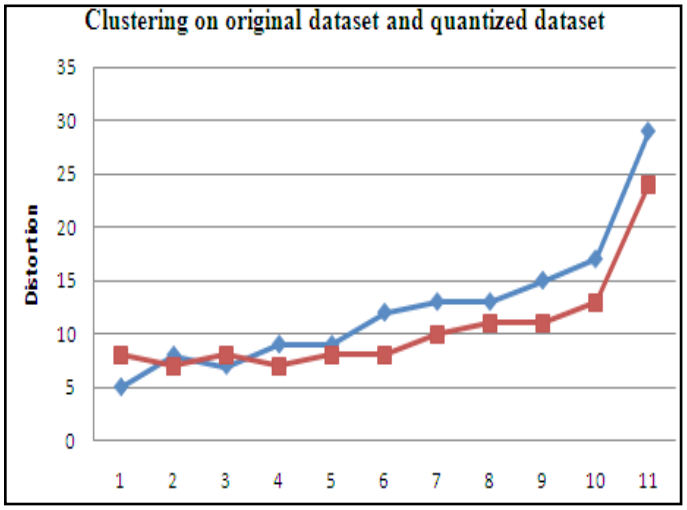

Fig. 4.5: Clustering on original data and quantized dataset for distortion using K-Means

The Vector Quantization techniques are efficiently applied in the development of speech recognition systems. In this paper, the proposed a novel vector quantization algorithm called K-meansLBG algorithm. It is used efficiently to increase the performance of the speech recognition system. The recognition accuracy obtained using K-meansLBG algorithm is better as compared to K-means and LBG algorithm. The average recognition accuracy of K-meansLBG algorithm is more than 2.55% using K-means algorithm while the average recognition accuracy of K-meansLBG algorithm is more than 1.41% using LBG algorithm.

V. 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. In this work, we have considered, for the first time, the issue of providing efficiency in privacy preserving mining. Our goal was to investigate the possibility of simultaneously achieving high privacy, accuracy and efficiency in the mining process. We first showed how the distortion process required for ensuring privacy can have a marked negative side-effect of hugely increasing mining runtime. Then, we presented our new K-MeansLBG algorithm that is specifically designed to minimize this side-effect through the application of symbol-specific distortion. We derived simple but effective formulas for estimating the performance beforehand and used optimization method to find the settings of the distortion parameters to get best privacy, accuracy and efficiency possible.

We also presented a simple but powerful optimization by which all additional counting incurred by privacy preserving mining is moved to the end of each pass over the database. Our experiments show that K-MeansLBG could simultaneously provide good

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

privacy, accuracy and efficiency. Specifically, less than 4 times slowdown with respect to Apriori in conjunction with 70-plus privacies and 90-plus accuracies, were achieved. In summary, K-MeansLBG takes a significant step towards making privacy-preserving mining of association rules a viable enterprise.

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.*