International Journal of Contemporary Research in Computer Science and Technology (IJCRCST) e-ISSN: 2395-5325
Volume1, Issue 9 (December '2015)

VECTOR QUANTIZATION FOR PRIVACY PRESERVING CLUSTERING IN DATA MINING

G.Satheesh,
Assistant Professor,
Department of Computer Science,
Mahendra Arts and Science College,
Kallipatti, Namakkal, Tamilnadu.

N.Suresh,
Assistant Professor,
Department of Computer Science,
Mahendra Arts and Science College,
Kallipatti, Namakkal, Tamilnadu.

Abstract: Huge volume of detailed personal data is regularly collected and sharing of these data is proved to be beneficial for data mining application. Such data include shopping habits, criminal records, medical history, credit records etc .On one hand such data is an important asset to business organization and governments for decision making by analyzing it .On the other hand privacy regulations and other privacy concerns may prevent data owners from sharing information for data analysis. 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 details 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.

Keyword: Data privacy, cluster, data mining, clustering, classification

I. INTRODUCTION

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. In this thesis, we provide the piecewise quantization approach for dealing with privacy preserving clustering.

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 [1]. Data mining is ready for application in the business community because it is supported by three technologies that are now sufficiently mature:

- Massive data collection
- Powerful multiprocessor computers
- Data mining algorithms

Data mining gets its name from the similarities between finding for important business information in a huge database — for example, getting linked products in gigabytes of stores canner data — and mining a mountain for a vein of valuable ore. These processes need either shifting through a large amount of material, or intelligently searching it to find exactly where the value resides. Data mining technology can produce new business opportunities by providing these features in databases of sufficient size and quality.

Prediction of trends and behaviors in automatic: The process of finding predictive information in large databases is automated by data mining. Questions that required extensive analysis traditionally can now be answered directly from the data — quickly with data mining technique. A typical example is targeted

marketing. It uses data on past promotional mailings to recognize the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of population likely to respond similarly to given events.

Discovery of previously unknown patterns: Data mining tools analyze databases and recognize previously hidden patterns in one step. The analysis of retail sales data to recognize seemingly unrelated products that are often purchased together is an example of pattern discovery. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying data that are anomalous that could represent data entry keying errors.

Data mining techniques can provide the features of automation on existing software and hardware platforms, and can be implemented on new systems as existing platforms are up graded and new products developed. On high performance parallel processing systems when data mining tools are used, they can analyze huge databases in minutes. Users can automatically experiment with more models to understand complex data by using faster processing facility. High speed make it possible for users to analyze huge quantities of data. Larger databases, in turn, yield improved predictions[2].

II. TECHNOLOGIES IN DATA MINING

The most commonly used techniques in data mining are:

Artificial neural networks: Non-linear predictive models that learn through training and resemble biological neural networks in structure.

Decision trees: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees(CART) and Chi Square Automatic Interaction Detection (CHAID).

Genetic algorithms: Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.

Nearest neighbor method: A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset. Sometimes called the k-nearest neighbor technique.

Rule induction: The extraction of useful if-then rules from data based on statistical significance.

III.APPLICATIONS OF DATA MINING

There is a rapidly growing body of successful applications in a wide range of areas as diverse as: analysis of organic compounds, automatic abstracting, credit card fraud detection, financial forecasting, medical diagnosis etc. Some examples of applications (potential or actual) are:

- a supermarket chain mines its customer transactions data to optimize targeting of high value customers
- a credit card company can use its data warehouse of customer transactions for fraud detection
- A major hotel chain can use survey databases to identify attributes of a 'high-value' prospect.

Applications can be divided into four main types:

- Classification
- Numerical prediction
- Association
- Clustering.

Data mining using labeled data (specially designated attribute) is called supervised learning. Classification and numerical prediction applications falls in supervised learning. Data mining which uses unlabeled data is termed as unsupervised learning and association and clustering falls in this category.

IV. DATA MINING AND PRIVACY

Data mining deals with large database which can contain sensitive information. It requires data preparation which can uncover information or patterns which may compromise confidentiality and privacy obligations. Advancement of efficient data mining technique has increased the disclosure risks of sensitive data. A common way for this to occur is through data aggregation. Data aggregation is when the data are accrued, possibly from various sources, and put together so that they can be analyzed. This is not data mining per se, but a result of the preparation of data before and for the purposes of the analysis. The threat to an individual's privacy comes into play when the data, once compiled, cause the data miner, or anyone who has access to the newly compiled data set, to be able to identify specific individuals, especially when originally the data were anonymous[3].

What data mining causes is social and ethical problem by revealing the data which should require privacy. Providing security to sensitive data against unauthorized access has been a long-term goal for the database security research community and for the government statistical agencies. Hence, the security issue has become, recently, a much more important area of research in data mining. Therefore, in recent years, privacy-preserving data mining has been studied extensively.

V.CLUSTERING

Division of data into groups of similar objects is called Clustering. Certain fine details are lost by representing the data by fewer clusters but it achieves simplification. It models data by its clusters. Data modeling puts clustering in a historical perspective rooted in mathematics, statistics, and numerical analysis. According to machine learning perspective clusters correspond to hidden patterns, the search for clusters is unsupervised learning and the resulting system represents a data concept. From a practical perspective clustering plays an important role in data mining applications such as scientific data exploration, information retrieval and text mining, spatial database applications, Web analysis, CRM, marketing, medical diagnostics, computational biology, and many others. Clustering can be shown with a simple graphical example:

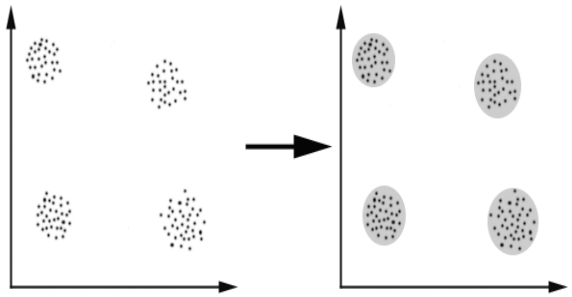

Figure 1: 4 clusters of data

In this case we easily identify the 4 clusters into which the data can be divided; the similarity criterion is distance*:* two or more objects belong to the same cluster if they are "close" according to a given distance(in this case geometrical distance).This is called distance-based clustering. Another kind of clustering is conceptual clustering: two or more objects belong to the same cluster if this one defines a concept common to all that objects. In other words, objects are grouped according to their fit to descriptive concepts, not according to simple similarity measures [4].

VI.CLASSIFICATION OF CLUSTERING ALGORITHMS

Categorization of clustering algorithms is not easy. In reality, groups given below overlap. For convenience we provide a classification as given in [1]. We are considering mainly hierarchical and partitioning methods.

Hierarchical Methods

- Agglomerative Algorithms
- Divisive Algorithms

Partitioning Methods

- Relocation Algorithms
- Probabilistic Clustering
- K-medoids Methods
- K-means Methods
- Density-Based Algorithms
- Density-Based Connectivity Clustering
- Density Functions Clustering

a) Hierarchical Clustering

Hierarchical clustering builds a cluster hierarchy or, in other words, a tree of clusters, also known as a dendrogram. Every cluster node contains child clusters; sibling clusters partition the points covered by their common parent. Such an approach allows exploring data on different levels of granularity. Hierarchical clustering methods are categorized into agglomerative (bottom-up) and divisive (top-down).An agglomerative clustering starts with one-point (singleton) clusters and recursively merges two or most appropriate clusters. A divisive clustering starts with one cluster of all data points and recursively splits the most appropriate cluster. The process continues until stopping criterion (frequently, the requested number k of clusters) is achieved.

Advantages of hierarchical clustering include:

- Embedded flexibility regarding the level of granularity
- Ease of handling of any forms of similarity or distance
- Consequently, applicability to any attribute types

Disadvantages of hierarchical clustering are related to:

- Vagueness of termination criteria
- The fact that most hierarchical algorithms do not revisit once constructed(intermediate) clusters with the purpose of their improvement

b) Partitioning Clustering

In contrast to hierarchical techniques, partitioned clustering techniques create a one level partitioning of the data points. If *K* is the desired number of clusters, then partitioned approaches typically find all *K* clusters at once. There are a number of partitioned techniques, but we shall only describe the K-means algorithm. It is based on the idea that a center point can represent a cluster. In particular, for K-means we use the notion of a centroid, which is the mean or median point of a group of points.

K-Means Methods- K-means clustering is a simple technique to group items into k clusters. There are many ways in which k clusters might potentially be formed. The quality of a set of clusters can be measured using the value of an objective function which is taken to be the sum of the squares of the distances of each point from the centroid of the cluster to which it is assigned. Its required that value of this function to be as small as possible.

Next k points are selected (generally corresponding to the location of k of the objects). These are treated as the centroids of k clusters, or to be more precise as the centroids of k potential clusters, which at present have no members. These points can be selected in any way, but the method may work better if *k* initial points are picked that are fairly far apart. Each of the points is assigned one by one to the cluster which has the nearest centroid. When all the objects have been assigned K clusters is formed based on the original k centroids but the 'centroids' will no longer be the true centroids of the clusters. Next centroids of the clusters are recalculated, and the previous steps are repeated, assigning each object to the cluster with the nearest centroid etc.

K means can be summarized as:

1. Choose a value of k
2. Select k objects in an arbitrary fashion. Use these as the initial set of k centroids.
3. Assign each of the objects to the cluster for which it is nearest to the centroid.
4. Recalculate the centroids of the k clusters.
5. Repeat steps 3 and 4 until the centroids no longer move.

Each object is placed in its closest cluster, and the cluster centers are then adjusted based on the data placement. This repeats until the positions stabilize. The results come in two forms: Assignment of entities to clusters, and the cluster centers themselves. The k-means algorithm also requires an initial assignment (approximation) for the values/positions of the k means. This is an important issue, as the choice of initial points determines the final solution.

VII. CLASSIFICATION OF PPDM

According to [4] work done in PPDM can be classified according to different categories. These
Are

Data Distribution- The PPDM algorithms can be first divided into two major categories, centralized and distributed data, based on the distribution of data. In a centralized database environment, data are all stored in a single database; while, in a distributed database environment, data are stored in different databases. Distributed data scenarios can be further classified into horizontal and vertical data distributions. Horizontal distributions refer to the cases where different records of the same data attributes are resided in different places. While in a vertical data distribution, different attributes of the same record of data are resided in different places. Earlier research has been predominately focused on dealing with privacy preservation in a centralized database. The difficulties of applying PPDM algorithms to a distributed database can be attributed to: first, the data owners have privacy concerns so they may not willing to release their own data for others; second, even if they are willing to share data, the communication cost between the sites is too expensive.

Hiding Purposes - The PPDM algorithms can be further classified into two types, data hiding and rule hiding, according to the purposes of hiding. Data hiding refers to the cases where the sensitive data from original database like identity, name, and address that can be linked, directly or indirectly, to an individual person are hided. In contrast, in rule hiding, the sensitive knowledge (rule) derived from original database after applying data mining algorithms is removed. Majority of the PPDM algorithms used data hiding techniques. Most PPDM algorithms hide sensitive patterns by modifying data.

Data Mining Tasks / Algorithms - Currently, the PPDM algorithms are mainly used on the tasks of classification, association rule and clustering. Association analysis involves the discovery of associated rules, showing attribute value and conditions that occur frequently in a given set of data. Classification is the process of finding a set of models (or functions) that describe and distinguish data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class

label is unknown. Clustering Analysis concerns the problem of decomposing or partitioning a data set (usually multivariate) into groups so that the points in one group are similar to each other and are as different as possible from the points another groups.

Privacy Preservation Techniques - PPDM algorithms can further be divided according to privacy preservation techniques used. Four techniques – sanitation, blocking, distort, and generalization – have been used to hide data items for a centralized data distribution. The idea behind data sanitation is to remove or modify items in a database to reduce the support of some frequently used item sets such that sensitive patterns cannot be mined. The blocking approach replaces certain attributes of the data with a question mark. In this regard, the minimum support and confidence level will be altered into a minimum interval. As long as the support and/or the confidence of a sensitive rule lie below the middle in these two ranges, the confidentiality of data is expected to be protected. Also known as data perturbation or data randomization, data distort protects privacy for individual data records through modification of its original data, in which the original distribution of the data is reconstructed from the randomized data. These techniques aim to design distortion methods after which the true value of any individual record is difficult to ascertain, but "global" properties of the data remain largely unchanged. Generalization transforms and replaces each record value with a corresponding generalized value.

VIII. TECHNIQUES OF PPDM

Most methods for privacy computations use some form of transformation on the data in order to perform the privacy preservation. Typically, such methods reduce the granularity of representation in order to reduce the privacy. This reduction in granularity results in some loss of effectiveness of data management or mining algorithms. This is the natural trade-off between information loss and privacy. Some examples of such technique as described in [5] are:

Randomization method - The randomization technique uses data distortion methods in order to create private representations of the records .In this which noise is added to the data in order to mask the attribute values of records In most cases, the individual records cannot be recovered, but only aggregate distributions can be recovered. These aggregate distributions can be used for data mining purposes. Data mining techniques can be developed in order to work with these aggregate distributions. Two kinds of perturbation are possible with the randomization method:

- Additive Perturbation - In this case, randomized noise is added to the data records. The overall data distributions can be recovered from the randomized records. Data mining and management algorithms re designed to work with these data distributions.
- Multiplicative Perturbation- In this case, the random projection or random rotation techniques are used in order to perturb the records.

The k-anonymity model and l-diversity-The *k*-anonymity model was developed because of the possibility of indirect identification of records from public databases. This is because combinations of record attributes can be used to exactly identify individual records. In the *k*-anonymity method, the granularity of data representation is reduced with the use of techniques such as generalization and suppression. This granularity is reduced sufficiently that any given record maps onto at least *k* other records in the data. The *l*-diversity model was designed to handle some weaknesses in the *k*-anonymity model since protecting identities to the level of *k*-individuals is not the same as protecting the corresponding sensitive values, especially when there is homogeneity of sensitive values within a group.

Distributed privacy preservation- In many cases, individual entities may wish to derive aggregate results from data sets which are partitioned across these entities. Such partitioning maybe horizontal(when the records are distributed across multiple entities) or vertical (when the attributes are distributed across multiple entities). While the individual entities may not desire to share their entire data sets, they may consent to limited information sharing with the use of variety of protocols. The overall effect of such methods is to maintain privacy for each individual entity, while deriving aggregate results over the entire data.

Downgrading Application Effectiveness - In many cases, even though the data may not be available, the output of applications such as association rule mining, classification or query processing may result in violations of privacy. This has lead to research in downgrading the effectiveness of applications by either data or application modifications.

IX. PRIVACY PRESERVING CLUSTERING: PROBLEM DEFINITION

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 [6]: Let *D* be 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, diseasediagnosis, 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 in [7].

X. VECTOR QUANTIZATION

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.

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.

Its further stated [7] a vector quantizer Q of dimension n and size s is a mapping: Q : R^n->c from a vector or a point in n-dimensional Euclidean space, R^n. ,, to a finite set C ={c1,c2, …Cs},the codebook, containing s output or reproduction points Ci ϵ R^n , called codeword. Associated with every vector quantizer with s codeword is a partition of R^n into s regions or cells Ri for iϵj = {1,2,…,s} where Rj = {xϵR^n: Q(X) = Ci} A distortion function is used to evaluate the overall quality degradation due to approximation of a vector by its closest representative from a codebook. For the mean-squared error distortion function d(x,Ci) between an input vector x and codeword Ci , an optimal mapping should satisfy two conditions: (a) the nearest neighbor condition and (b) the centroid condition.

Nearest neighbor Condition
For a given codebook, the optimal partition R = {Ri : i=1,…,s} satisfies:

Ri = {x: $d(x,c_i) < d(x,c_i)$;Vj}where c_i is the codeword representing partition Ri . Given a point *x* in the dataset, the encoding function for *x*,

Encoding(*x*) = Ci only if $d(x,c_i) < d(x,c_i)$;Vj

Centroid condition

For a given partition region Ri (I = 1,….,s) , the optimal reconstruction vector (codeword) satisfies: Ci = centroid(Ri) where the centroid of a set *R* = {*xi*: *i* = 1,…|*R*|} is defined as:

$$centroid(R) = \frac{1}{|R|}\sum_{i=1}^{|R|} x_i$$

These two conditions are used by an iterative procedure, the generalized Lloyd algorithm and this is used in k means.

CONCLUSION

This paper gives a detail about vector quantization for privacy preserving clustering in data mining. In this paper, we have discussed about data mining applications, techniques and privacy policies. We describe the Clustering and its types of techniques, Privacy-Preserving Clustering with vector quantization approach.

REFERENCES

[1] Berkhin Pavel, A Survey of Clustering Data Mining Techniques, Springer Berlin Heidelberg, 2006.

[2] Agrawal R., Srikant R. Privacy preserving data mining. In: Proceeedings of the ACM SIGMOD Conference of Management of Data, pp. 439–450. ACM (2000).

[3] Bramer Max, Principles of Data Mining, London, Springer, 2007.

[4] Wu Xiaodan, Chu Chao-Hsien, Wang Yunfeng, Liu Fengli, Yue Dianmin, Privacy Preserving Data Mining Research: Current Status and Key Issues, Computional Science ICCS 2007,4489(2007), 762-772.

[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/

