Privacy Preserving Clustering In Data Mining

A thesis submitted in partial fulfillment of the requirements for the degree of

Bachelor of Technology In Computer Science and Engineering

By

Jitendra Kumar

Roll No: 10606049

&

Binit Kumar Sinha

Roll No: 10606051

Department of Computer Science and Engineering

National Institute of Technology Rourkela

Rourkela - 769008

Privacy Preserving Clustering
In Data Mining

A thesis submitted in partial fulfillment of the requirements for the degree of

Bachelor of Technology In Computer Science and Engineering

By

Jitendra Kumar

Roll No: 10606049

&

Binit Kumar Sinha

Roll No: 10606051

Under the guidance of

Prof. S.K. Jena

Department of Computer Science and Engineering

National Institute of Technology Rourkela

Rourkela - 769008

National Institute of Technology

Rourkela

CERTIFICATE

This is to certify that the thesis entitled, "Privacy Preserving Clustering in Data Mining" submitted by Jitendra Kumar and Binit Kumar Sinha in partial fulfillments for the requirements for the award of Bachelor of Technology Degree in Computer Science Engineering at National Institute of Technology, Rourkela is an authentic work carried out by them under my supervision and guidance.

To the best of my knowledge, the matter embodied in the thesis has not been submitted to any other University / Institute for the award of any Degree or Diploma.

Date:

Prof. S. K. Jena
Dept. of Computer Science Engineering
National Institute of Technology
Rourkela - 769008

ACKNOWLEDGEMENT

On the submission of my Thesis report, we would like to extend our gratitude & sincere thanks to our supervisor Prof. S.K.Jena, Professor, Department of Computer Science and Engineering, NIT Rourkela for his constant motivation and support during the course of my work in the last one year. I truly appreciate and value his esteemed guidance and encouragement from the beginning to the end of this thesis. He has been my source of inspiration throughout the thesis work. A special acknowledgement goes to Prof. Korra Sathya Babu for extending his support during entire duration of the project and giving us insights into the subject matter. We would also like to convey our sincerest gratitude and indebtedness to all other faculty members and staff of the Department of Computer Science and Engineering, NIT Rourkela, who bestowed their great effort and guidance at appropriate times without which it would have been very difficult on our part to finish the project work.

Jitendra Kumar
Roll No: 10606049

Binit Kumar Sinha
Roll No: 10606051

Contents

List of Tables

List of figures

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 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.

Chapter 1

Introduction

1.1 Data mining
1.2 Scope of data mining
1.3 Technologies in data mining
1.4 Application of data mining
1.5 Data mining and privacy

Over the last twenty years, there has been a 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 its 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.

1.1Data mining

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. 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

1.2 Scope of data mining

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 store

scanner data — and mining a mountain for a vein of valuable ore. These processes need either shifting through an 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,

· *Automated prediction of trends and behaviors.* 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 a population likely to respond similarly to given events.

· *Automated 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 upgraded 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.

1.3 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.

1.4 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.

1.5 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.

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.

We will further see the research done in privacy area .In chapter 3 general survey of privacy preserving methods used in data mining is presented.

Chapter 2

Clustering and its classification

2.1 What is clustering?

2.2 Classification of Clustering Algorithms

2.2.1 Hierarchical Clustering

2.2.2 Partitioning Clustering

2.1 What is 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:

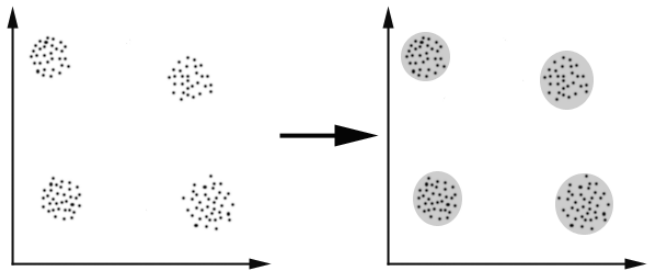


Figure 1: Example showing 4 cluster of data

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.

2.2Classification 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

2.2.1Hierarchical 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 a 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

2.2.2 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.

As explained in [3] the entire algorithm of 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.

Chapter 3

Survey on Privacy Preserving Data Mining (PPDM)

3.1 Classification of PPDM

3.2 Techniques of PPDM

3.1 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 in other 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.

3.2 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 may be 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 a 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.

Chapter 4

Privacy Preserving Clustering: Problem and its solution

4.1 Privacy preserving clustering: Problem definition

4.2 Piecewise Vector Quantization as a solution for Privacy preserving clustering

4.1 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 as in [6]: 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 in [7].
This chapter covers this method in detail.

4.2 Piecewise Vector Quantization as a solution for Privacy preserving clustering

What is 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 in [7] a vector quantizer Q of dimension n and size s is a mapping: $Q: R^n \rightarrow C$ from a vector or a point in n-dimensional Euclidean space, R^n, to a finite set $C = \{c_1, c_2, \dots, c_s\}$,the codebook, containing s output or reproduction points $C_i \epsilon R^n$, called codewords. Associated with every vector quantizer with s codewords is a partition of R^n into s regions or cells R_i for $i \epsilon j \equiv \{1,2, \dots, s\}$ where$R_j = \{x \epsilon R^n : Q(x) = C_i\}$. 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 $d(x, c_i)$ between an input vector x and a codeword c_i , 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 $R = \{R_i : i = 1, \dots, s\}$satisfies:

$$R_i = \{x: d(x, c_i) \leq d(x, c_j); \forall j\}$$

where C_i is the codeword representing partition R_i. Given a point *x* in the dataset, the encoding function for *x*,

Encoding(*x*) = $C_i$only if $d(x, c_i) \leq d(x, c_j)\ \forall j$.

• *Centroid condition*

For a given partition region R_i (i = 1,…,*s*), the optimal reconstruction vector (codeword) satisfies: C_i = centroid(R_i) 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.

Piecewise Vector Quantization

We tried to preserve privacy of data by applying VQ on segments of the datasets and transforming datasets into new datasets. This method segmentize every point (row) of datasets into w segments each of length L. Each segment is represented by the closest, based on distance measure, entry in the codebook(containing code vector of L dimension).The point(row) of data, which is segmentize , is now represented by new transformed row of data formed by joining the new segments formed.

In order to create the codebook, each point(row) data in a training set is partitioned into *w* segments, each of length *l*. These segments form the samples used to train the codebook using K means clustering method. Codebook contains the centroid of all clusters formed through K means clustering method. Each segment of length L is approximated and represented by centroid , of length L, of cluster in which it falls as resulted by K means clustering method i.e. closest based on distance measure in codebook .These new w segments formed are joined to form a new point(row) which is replacement of original point(row) of dataset to which this w segments belong. Thus preserving privacy as dataset is completely transformed to new dataset. It also

gives accuracy in terms of clusters as similarity between the points in datasets are preserved .Similarity depends on Euclidean distance between points which is more or less preserved.

Codebook Generation and transformation

In order to build codebook and transformation we use K means clustering method. Each point(row) of datasets is segmentize into w segment of length L which in turn decomposes original datasets into w datasets of dimension L. Each of the w datasets is processed through K means clustering method which forms k cluster of each dataset. The points falling in Cluster C_i is represented by the centroid(C_i) of that cluster. Centroid(C_i) is also of L length. Thus Codebook consist of centroid of all cluster. These w datasets are further joined to form transformed datasets of original such that all w segment of original point(row) of data which are now approximated as centroid of cluster are joined.

The K means starts with an initial codebook and repeats the iteration in which the two conditions for optimality are used sequentially, i.e., given a codebook (created in the previous iteration), the optional partition of the space R^n is found using the nearest neighbor condition, and given a partitioning, the codebook is updated using the centroid condition. The distortion function between each vector and its nearest codeword is then calculated. This process repeats until the fractional drop of the distortion between two consecutive iterations becomes smaller than a given threshold.

The quality of the codebook obviously affects the performance of a vector quantizer, thus it is important to obtain a good codebook. During the training process, the codebook keeps being updated until a certain optimization condition is satisfied. This condition is often the minimization of a distortion measurement $d(x, c_i)$ (e.g., the mean squared error) between a training sample x and its corresponding codeword c_i, or the fractional drop of the distortion becoming less than a given threshold. Usually the quality of a codebook is considered poor when the average of $d(x, c_i)$ is above a given threshold. Although it is possible to reduce the expected distortion by increasing the number of codewords, the computation and storage expenses will also increase at the same time. So, there is a tradeoff between quality and cost.

Next we discuss the steps involved in our method

Steps involved in Piecewise Vector Quantization

To address our privacy preserving clustering technique we define the following six steps to be followed:

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 $X = x_1\ x_2\ x_3 \ldots\ldots x_n$. Dataset contain "m" row of data.

In our case we use Water Treatment dataset available at UCI Repository.

Step 2: Segmentation – Dataset is segmentize into w datasets by decomposing each row of data into w segments each of length L. Let each row of data is represented by $X = x_1\ x_2\ x_3 \ldots\ldots x_n$ .Then it is segmentize into w segments as given below.

1st Segment $Y_1 = x_1\ x_2\ x_3 \ldots\ldots\ldots x_L$
2nd Segment $Y_2 = x_{L+1}\ x_{L+2}\ x_{L+3} \ldots\ldots\ldots x_{2L}$
3rd Segment $Y_3 = x_{2L+1}\ x_{2L+2}\ x_{2L+3} \ldots\ldots\ldots x_{3L}$
.
.
.
With Segment $Y_w = x_{(w-1)L+1}\ x_{(w-1)L+2}\ x_{(w-1)L+3} \ldots\ldots\ldots x_{wL}$

Dataset D is decomposed into $D_1\ D_2\ D_3$ .....D_W 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 step5.

Step 3: Clustering for Codebook Generation – In order to generate codebook which is helpful in data transformation, K means clustering algorithm is used. This algorithm takes Euclidean distance as the similarity measure. Euclidean distance between two data point X_i and X_j is given as $d(i,j)$

$$d(i,j) = \left[\sum_{k=1}^{n} |x_{ik} - x_{jk}|^2 \right]^{1/2}$$

Clustering is performed on each dataset, resulting on account of decomposition of original dataset D. Each dataset is divided into k regions called clusters and each cluster is represented by its centroid. Every single row data falls in exactly one cluster.

Let dataset D_i is divided into k clusters

$$C_{i1}\ C_{i2}\ C_{i3}\ \ldots\ldots\ldots C_{iK}$$

And each cluster is represented by its centroid ...i.e. centroid of j^{th} cluster of i^{th} dataset is:

$$\text{Centroid } (C_{ij}) = \{\ c_1\ c_2\ c_3\ c_4 \ldots\ldots\ldots c_L\}$$

This centroid is also of length L. Codebook for each dataset consist of centroid of all k cluster.

$$\text{Codebook} = (\ \text{centroid}(C_{i1}))\ ,\ \text{centroid}(C_{i2}\ ),\ \text{centroid}(C_{i3}),\ldots\ldots\ldots\text{centroid}(C_{i4}))$$

Step 4: Data Transformation by Quantization – Each decomposed dataset D_i is transformed into new dataset D_i' 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.

Let row data be Y_i…i.e. i^{th} segment of X in step2 that fall in D_i

$$Y_i = x_{(i-1)L+1}\quad x_{(i-1)L+2}\quad x_{(i-1)L+3}\quad \ldots\ldots\quad x_{iL}$$

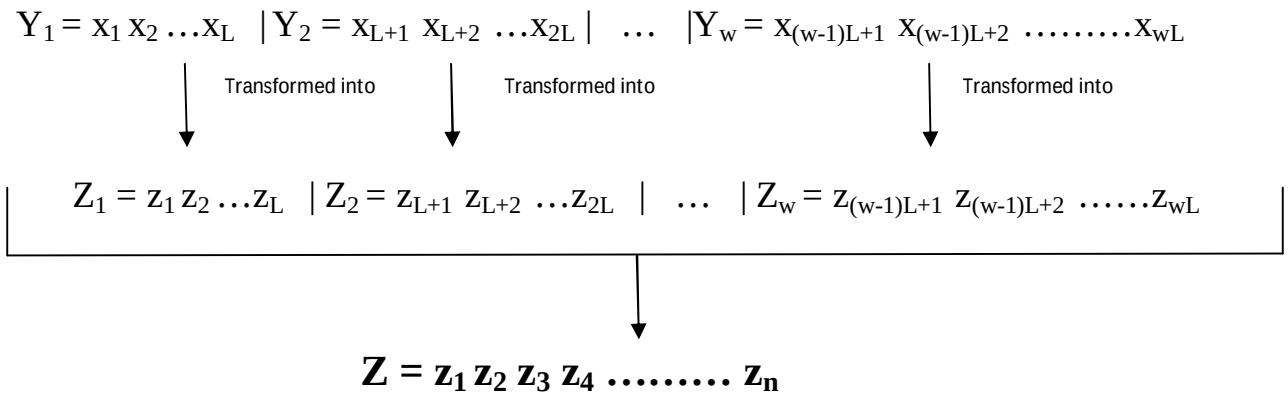
And it falls in cluster j^{th} cluster so this Y_i will be replaced by Z_i = Centroid (C_{ij}) i.e. centroid of j^{th} cluster of i^{th} dataset as this is the point which is nearest in code book.

$$Y_i \xrightarrow{\text{transformed into}} Z_i$$

The transformation function is:

$$T(Y_i) = \{\text{Centroid}(C_j)\text{: only if } d(Y_i\ , \text{centroid}(C_j)) <\ d(Y_i, \text{centroid}(C_z)) \text{ for } \forall z\ \}$$

Step 5: Reformation of Dataset – Each segment Y_i of row data X formed as segmentation step in step 2 is transformed into Z_i 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 dimension transformed row data which replaces the X in the original dataset.



This is repeated for each row and a complete transformed row data is formed for each. All transformed data set is formed to form new Dataset. This way perturbed dataset $D^{/}$ is formed from original dataset D.

Step 6 – Comparison for accuracy and 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.

Performance evaluation based on two metric that is accuracy and distortion metric. Both are discussed in detail below.

1. Accuracy

Our evaluation approach focuses on the overall quality of generated clusters We compare how closely each cluster in the modified dataset matches its corresponding cluster in the original dataset. This method was used for checking accuracy in method DRBT in [6].As stated in [6] to do so, we first identify the matching of cluster by computing the matrix of frequencies showed in below table. We refer to such a matrix as the clustering membership matrix (CMM), where the rows represent the clusters in the original dataset, the columns represent the clusters in the transformed dataset, and $Freq_{i,j}$is the number of points in cluster C_i that falls in cluster C_j' in the transformed dataset. After computing the frequencies $Freq_{i,j}$, we scan the CMM calculating precision, recall, and F-measure for each cluster C_j' with respect to C_i in the original dataset. These formulas are given by the following equations:

	C'_1	C'_2	...	C'_k
C_1	$Freq_{1,1}$	$Freq_{1,2}$	...	$Freq_{1,k}$
C_2	$Freq_{2,1}$	$Freq_{2,2}$	...	$Freq_{2,k}$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
C_k	$Freq_{k,1}$	$Freq_{k,2}$	...	$Freq_{k,k}$

Table 1: Showing number of points in cluster C_i that falls in cluster C_j' in the transformed dataset

$$Precision(P) = \frac{Freq_{i,j}}{|C_j'|}$$

$$Recall(R) = \frac{Freq_{i,j}}{|C_i|}$$

$$F - measure(F) = \frac{2 \times P \times R}{(P+R)}$$

Where |X| is the number of points in the cluster X.

For each cluster C_i , we first find the cluster C_j' that has the highest F-measure among all the C_L' , $1 \leq L \leq k$. Let $F(C_i)$ be the highest F-measure for the cluster C_i , we denote the overall F-measure (OF) as the weighted average of $F(C_i)$,$1 \leq i \leq K$, as follows:

$$OF = \frac{\sum_{i=1}^{k} |c_i| \times F(c_i)}{\sum_{i=1}^{k} |c_i|}$$

2. Distortion

Distortion is measured with the help of comparison between Original dataset and modified dataset. Each tuple X_i in original dataset , of m rows and each row is of n attributes , which is transformed into Y_i in modified dataset is used to compute distortion in that tuple by calculating dissimilarity between them through Euclidean distance by the formula.

$$D(Xi, Yi) = [\sum_{k=1}^{n} |Xik - Yik|^{\frac{1}{2}}]^2$$

The overall distortion is computed by the following equation on full datasets

$$Distortion = \frac{1}{mn} \sum_{i=1}^{m} [\sum_{k=1}^{n} |Xik - Yik|^{\frac{1}{2}}]^2$$

Chapter 5

Experiment and Results

5.1 Dataset Information

5.2 Methodology

5.3 Result

5.1 Dataset Information

Water treatment [8] dataset available on UCI Machine Learning Repository was taken for experimenting. It consists of 527 records and 38 attributes. Attributes type is integer or real.

Dataset	Water Treatment
No .of records	527
No. of attributes	38

Table 2 : Dataset Information

5.2 Methodology

Series of experiment was performed varying segment size(L) i.e. segment number (w) varies and varying number of cluster for quantization(K). Our evaluation approach focused on the overall quality of generated clusters after transforming dataset and the distortion produced in the dataset. Experiment was based on following steps

1> We modified the dataset by dealing with the missing value .To do so we replace it with average value of that attribute over the whole dataset.

2> We applied piecewise vector quantization method to transform dataset.

3> We selected K means to find the clusters in our performance evaluation. Our selection was influenced by following aspects (a) K-means is one of the best known clustering algorithm and is scalable. (b) K-means was also used in our codebook generation step. Number of cluster to be find from original and transformed dataset was taken same as number of cluster for quantization. This is the limitation in our experiment, experiment can also be used to find result using two different value, one for number of cluster for quantization(K) and other for number of cluster to be find from original and transformed dataset. Although we performed experiment taking same value.

4> We compared how closely each cluster in the transformed dataset matches its corresponding cluster in the original dataset. We expressed the quality of the generated clusters by computing the F-measure.

5> We compared the distortion produced due to transformation of dataset by the distortion metric.

5.3 Result:

Experiment was performed for measuring distortion in transformed dataset on different K value keeping the L constant and on different L value keeping the K constant. The result which came from our experiment is shown in Table 2 and corresponding graph between distortion and K and between distortion and L value is drawn as shown next pages.

L/K	**5**	**10**	**15**	**20**	**25**	**30**	**35**	**40**	**45**	**50**
2	42.82	22.85	14.08	11.81	11.05	11.06	10.25	9.033	8.50	7.959
3	42.84	22.89	14.18	11.89	11.16	11.12	10.34	9.15	8.62	8.06
4	43.02	23.172	14.87	12.31	11.57	11.58	10.83	9.68	9.14	8.43
5	43.36	23.46	14.97	12.79	12.25	11.99	11.43	10.26	9.80	9.34
6	43.25	23.56	15.21	13.40	12.76	12.48	11.59	10.37	10.06	9.44
7	43.26	23.60	15.24	13.42	12.78	12.53	11.62	10.36	10.08	9.74
8	43.28	23.72	15.34	13.57	12.94	12.63	11.76	10.43	10.24	9.89
9	44.56	25.53	17.94	16.18	15.37	13.32	14.33	12.51	13.12	12.16
10	44.61	25.61	18.05	16.28	15.45	15.39	14.43	12.90	13.17	12.22
11	44.55	25.68	18.11	16.46	15.61	15.50	14.48	13.03	12.60	12.57
12	44.557	25.73	18.20	16.59	15.75	15.66	14.42	13.38	13.47	12.68
13	44.558	25.70	18.21	16.597	15.71	15.37	14.43	13.45	13.43	12.59
14	44.558	25.705	18.22	16.59	15.70	15.39	14.43	13.40	13.45	12.592
15	45.63	27.22	21.33	18.04	17.34	16.80	15.93	14.564	14.14	14.022
16	45.635	27.221	21.337	18.046	17.36	16.81	15.93	14.567	14.60	14.03
17	45.638	27.227	21.34	18.043	17.31	16.77	16.01	14.28	14.63	14.05
18	45.648	27.298	21.35	18.49	17.34	16.97	16.12	14.69	14.76	14.11
19	45.649	27.242	21.36	18.402	17.20	16.92	16.21	14..70	14.77	14.16

Table 3 : Distortion value at different K and L value

Figure 2 shows Distortion Vs Segment Size(L). Segment size varies from 2 to 19 and its corresponding distortion measured by distortion metric on various value of K is shown. It can be easily concluded that distortion increases with increase in L and it's obvious as more the value of L more the attribute is affecting for quantization so more is the irregularity and more the distortion.

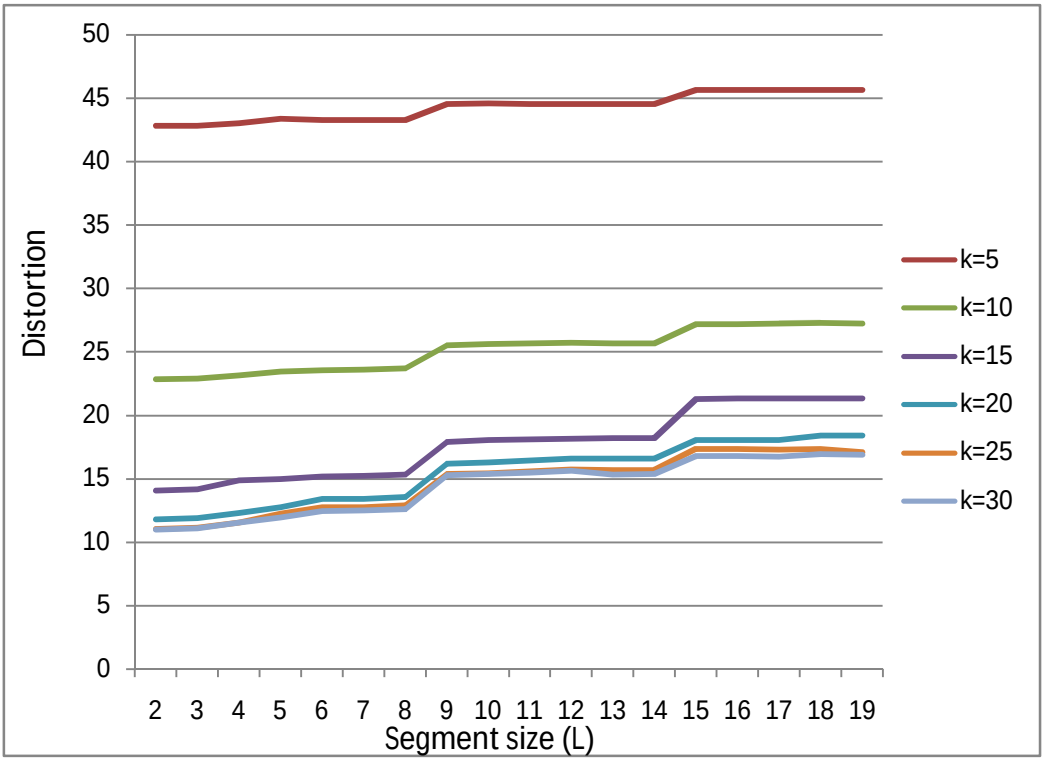


Figure 2 : Line Graph of Distortion Vs Segment size(L) at different K value

Distortion also reduces with increase in K as with increase in K less number of row points are used for quantization (as average numbers of points per cluster reduces and codebook generation ,that is later used for quantization, take place as mean of all points falling in a cluster) so less is the distortion. It's also shown in Figure 3 as a line graph and Figure 4 as bar graph between Distortion and number of cluster for quantization(K) at various segment size values, Distortion leads to loss of information which can leads to loss in information in cluster. So it should be reduced.

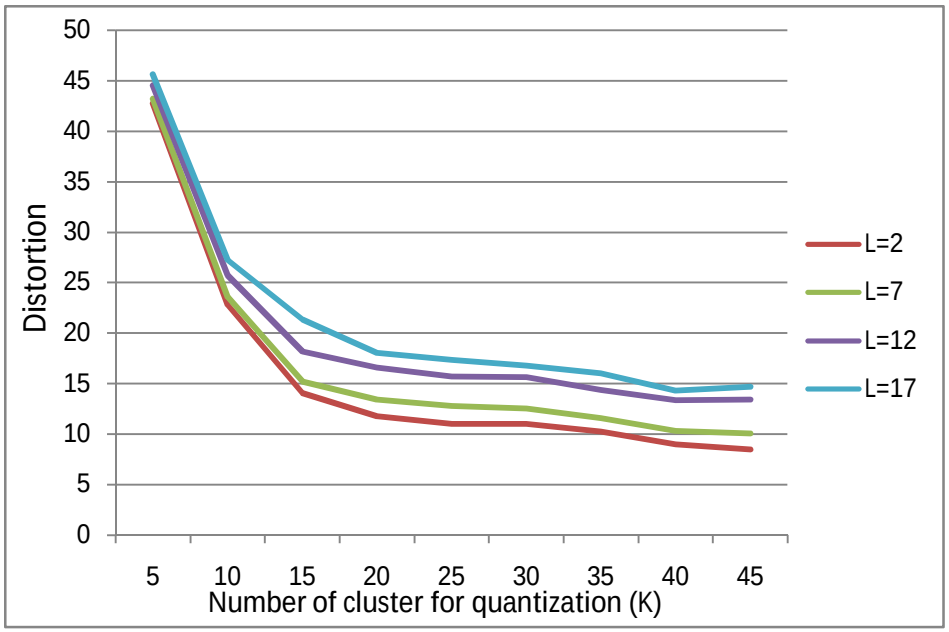


Figure 3 : Line Graph of Distortion Vs Number of cluster for quantization(K) at different L

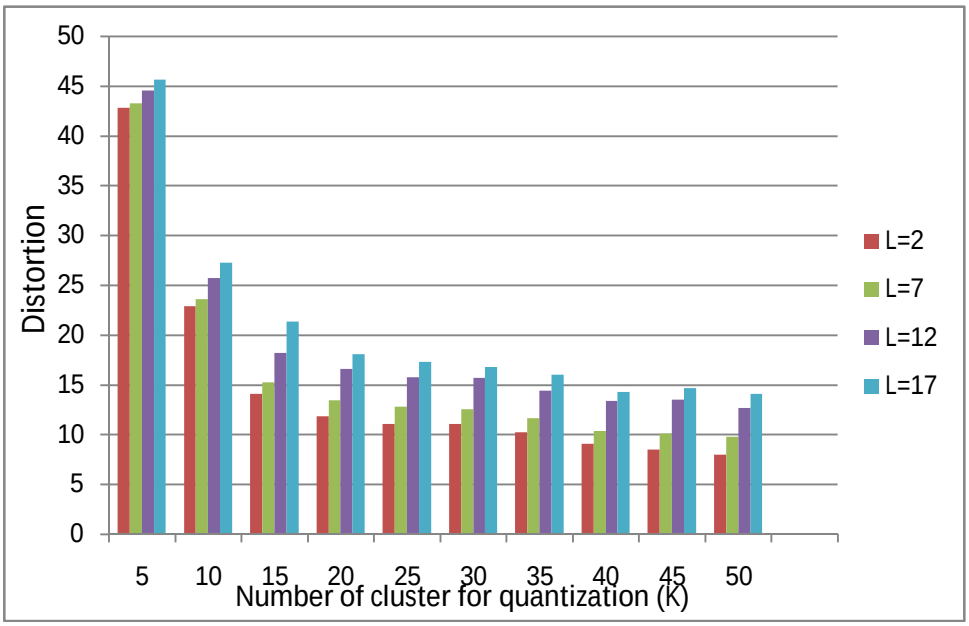


Figure 4 : Column Graph of Distortion Vs Number of cluster for quantization(K) at different L

Experiment was performed for measuring Fmeasure in transformed dataset on different K value keeping the L constant and on different L value keeping the K constant. The result which came from our experiment is shown in Table 3 and corresponding graph between distortion and K and between distortion and L value is drawn as shown next pages.

L/K	5	10	15	20	25	30	35	40	45	50
2	0.84	0.77	0.72	0.69	0.73	0.74	0.609	0.67	0.74	0.77
3	0.849	0.779	0.747	0.707	0.74	0.744	0.641	0.65	0.77	0.779
4	0.849	0.813	0.748	0.742	0.745	0.757	0.646	0.69	0.76	0.785
5	0.851	0.785	0.702	0.719	0.756	0.756	0.654	0.699	0.707	0.80
6	0.849	0.822	0.682	0.740	0.76	0.761	0.675	0.663	0.70	0.760
7	0.849	0.80	0.69	0.71	0.735	0.748	0.66	0.68	0.69	0.79
8	0.849	0.789	0.706	0.7125	0.753	0.757	0.617	0.68	0.69	0.80
9	0.851	0.822	0.701	0.687	0.79	0.808	0.658	0.652	0.699	0.783
10	0.846	0.813	0.738	0.675	0.76	0.77	0.618	0.72	0.721	0.83
11	0.848	0.820	0.730	0.675	0.728	0.732	0.624	0.68	0.688	0.76
12	0.849	0.806	0.72	0.67	0.77	0.781	0.67	0.70	0.704	0.78
13	0.848	0.781	0.707	0.705	0.723	0.729	0.645	0.718	0.699	0.81
14	0.848	0.823	0.740	0.691	0.747	0.755	0.642	0.730	0.697	0.760
15	1.0	0.80	0.784	0.673	0.78	0.767	0.608	0.671	0.741	0.797
16	1.0	0.80	0.71	0.675	0.75	0.768	0.635	0.708	0.754	0.80
17	1.0	0.82	0.73	0.660	0.781	0.80	0.672	0.71	0.717	0.78
18	1.0	0.81	0.786	0.673	0.791	0.744	0.644	0.719	0.742	0.80
19	1.0	0.79	0.783	0.647	0.789	0.79	0.66	0.711	0.747	0.77

Table 4 : Fmeasure value at different K and L value

Figure 5 shows Fmeasure calculated at different value ok K (number of cluster that is used for quantization) .It has been found that initially at low K value Fmeasure is high but slowly it decreases with increase in value of K .Again at certain point it again rises in this graph its approx

at K=25, after rising for certain value it reaches to local maxima here in graph is at K=30 from whereon it again decreases with some K value. Finally the graph rises with increasing K value till 1 as when K=527 that is equal to number of row in dataset then there is no transformation in dataset and Fmeasure is 1.For some L value there is different K value for local maxima but mostly in our experiment it is K=30.Here below is explanation of this behavior of graph from our perspective and reason why we took K=30 later for water treatment dataset for transforming dataset.

1> It's giving high Fmeasure at low value of K because there is little scope of point to jump from one cluster to other due to low K value. And as shown by Figure 3 distortion is large at low K value so information loss is more .Whereas it gives high privacy as number of points used for quantization is more but there is high loss of information so these K value cannot be used. At extreme extent if we take k=1 then every point will be quantizes to a same point and all falls in single cluster and as we took number of cluster to be formed from transformed dataset same as K value so Fmeasure is also high. Instead of taking same value as K for calculation of Fmeasure if we have taken other value like 2 then Fmeasure will be worse in this situation so keeping this in mind we rejected the low K value.

2> It's giving high Fmeasure at high value of K because points are somewhat approximately quantized as less number of points are being used for quantization so there is little chance of points to jump from one cluster to other due to points being approximately same as original. Due to points being approximately same or very nearer to original distortion is also less. This in turn also effects privacy as points can easily be approximated. At extreme extent if K=527 that is equal to number of row in dataset then there is no transformation in dataset and Fmeasure is 1.

3> So due to above two explanation it is required that there should be some optimum value ok K here its K=30 where Fmeasure is also high and distortion is also not very high or low effecting privacy.

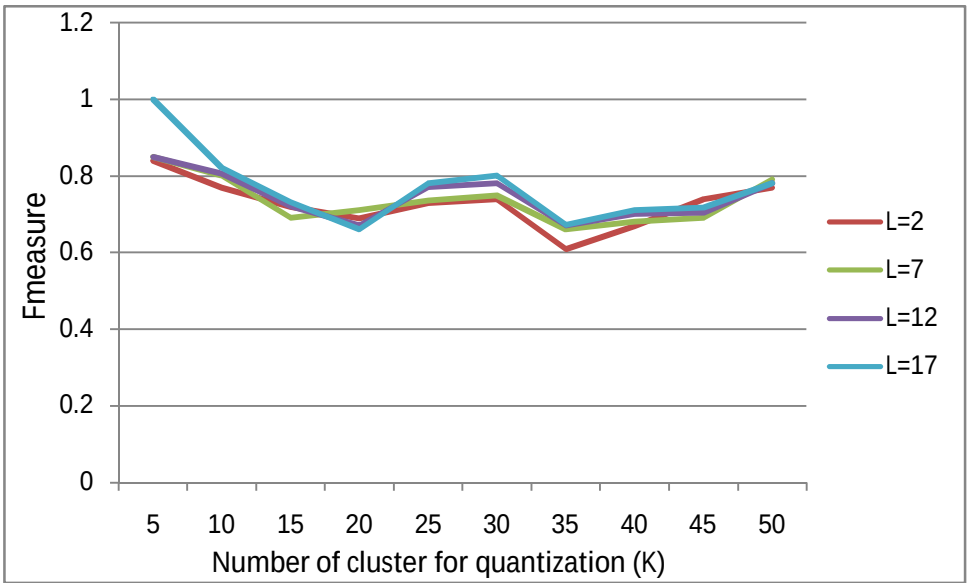


Figure 5 : Line Graph of Fmeasure Vs Number of cluster for quantization(K) at different L

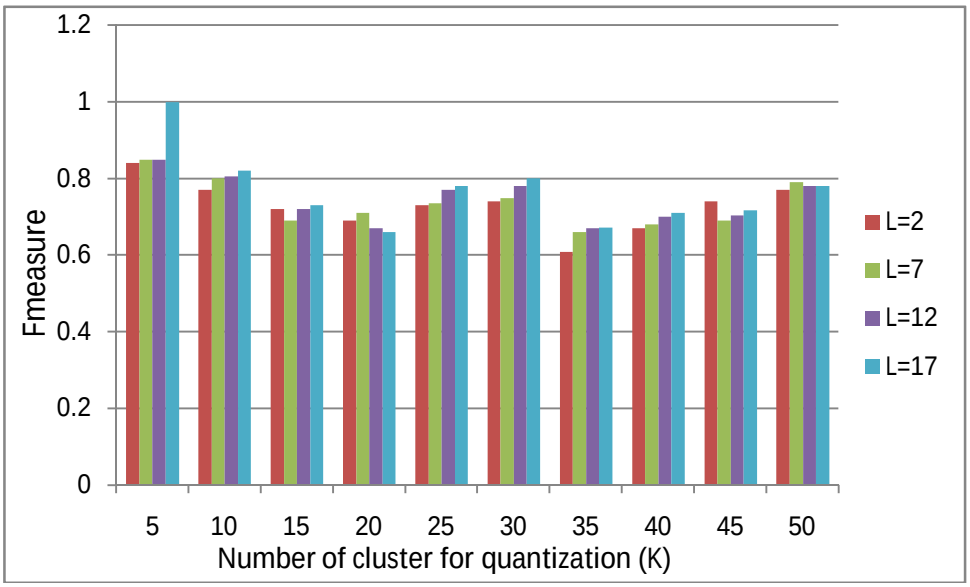


Figure 6 : Column Graph of Fmeasure Vs Number of cluster for quantization(K) at different L

Figure 7 shows the trend between segment size(L) and Fmeasure at different value of K. There was no general trend found like it increases with L or decreases. At some point in graph Fmeasure increases at other it decreases.

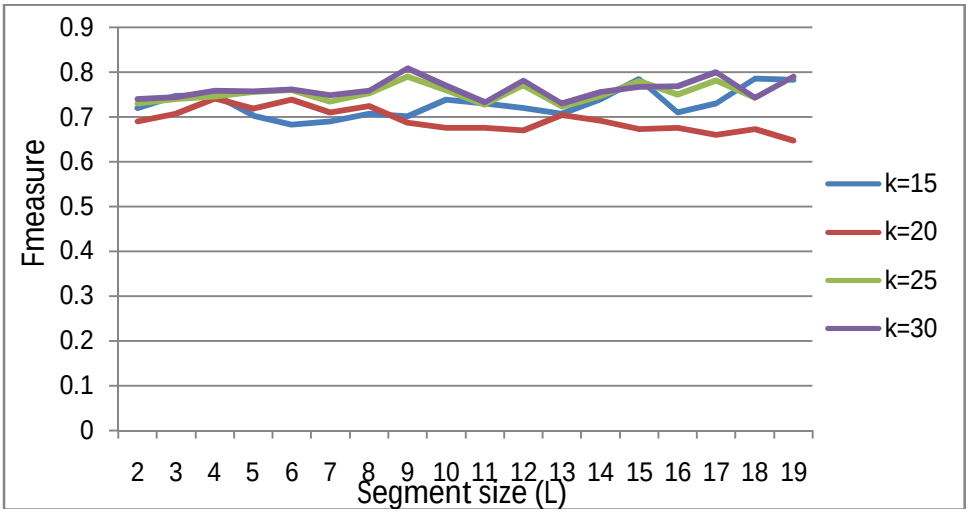


Figure 7 : Line Graph of Fmeasure Vs Segment size(L) at different K value

With the previous analysis we took K=30 as numbers of cluster for quantization. Now Figure 8 shows the value of Fmeasure at K=30 with varying segment size(L). It can be seen from graph that Fmeasure is highest at L=9. So we found that for water treatment dataset for privacy preserving its giving best result for K=30 and L=9 with Fmeasure as 0.808 and distortion as 13.32 and thus preserving privacy due to distortion.

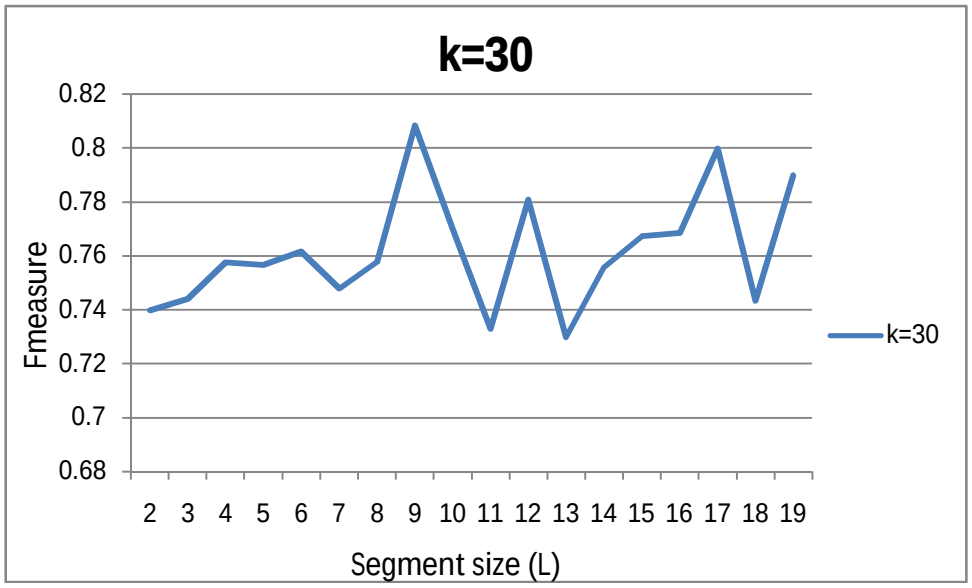


Figure 8 : Line Graph of Fmeasure Vs Segment size(L) at K=30 value

Chapter 6

Conclusion and Future Work

6.1 Conclusion

6.2 Future work

6.1 Conclusion

This thesis gives a different approach of using piecewise vector quantization for privacy preserving clustering in data mining. In this thesis, we have showed analytically and experimentally that Privacy-Preserving Clustering is to some extent possible using piecewise vector quantization approach. To support our claim we used water treatment dataset available on UCI Machine Learning Repository and performed experiment on it varying segment size and number of cluster for quantization. We evaluated our method taking into account two important issues: distortion and Fmeasure (quality of the clustering results).

Our experiment showed the variation of Fmeasure and distortion with segment size and number of cluster for quantization. It was found optimum segment size and number of cluster for quantization which give nice Fmeasure and distortion and in turn privacy for water treatment dataset. Our experiments revealed that using piecewise quantization approach a data owner can meet privacy requirements without much losing the benefit of clustering since the similarity between data points is preserved or marginally changed.

6.2 Future work

As future work new and effective quantization method can be used rather than K means approach that we have used. K nearest neighbor approach is one of the approach which can give better result. Moreover instead of segmenting sequence wise i.e. the first five attributes fall in single segment next five in other, the attributes which are related in some way can be kept in a single segment in future this can give good quantization as attributes play a role in quantization of other attribute. Module handling outliers can also be incorporated in this method to give good result.

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/