ISSN: 2581-3404 (Online)
International Journal of Innovative Research in Technology and Management (IJIRTM), Volume-3, Issue-6, 2019

# Clustering in Data Mining Using Piecewise Vector Quantization Approach

**Rochak Mahato[1], Prof. Gagan Sharma[2]**

**[1]M. Tech. Scholar, [2]Asst. Professor, Department of Computer Science & Engineering**

**[1,2]Sri Satya Sai College of Engineering Bhopal (M.P)**

**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.

**Keywords:-** Data Mining (DM), Knowledge, classification, Learning Analytics (LA), Water Treatment database, F-measure.

## 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 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.

## II Data 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

## III 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 storescanner 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.

## IV 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.

## V Literature Review

Many works have been carry out to explore the benefits of using Piecewise Vector Quantization Approach. The work done by various authors describe below:-

Md. Hedayetul Islam Shovon, Mahfuza Haque, "An Approach of Improving Student's Academic Performance by using K-means clustering algorithm and Decision tree", 2018 , In this study we make use of data mining process in student's database using k-means clustering algorithm and decision tree technique to predict student's learning activities. We hope that the information generated after the implementation of data mining and data clustering technique may be helpful for instructor as well as for students. This work may improve student's performance; reduce failing ratio by taking appropriate steps at right time to improve the quality of education.

Dr Priyanka Sharma, "Performance Prediction Of Students Using Distributed Data Mining" 2017 Distributing training and testing task of classification on each Node and central Node respectively, improves classification and prediction task on large and distributed data. Predicting student performance is useful to create efficient and good quality student work force ,by predicting student at risk and give them better training to improve their performance will surely beneficial for their individual results and also for academic institution profile. At present prediction of student results of engineering students applying this application on students of different areas.

M.I. López, J.M Luna, C. Romero, S. Ventura, "Classification via clustering for predicting final marks based on student participation in forums" 2012. the result obtained, the experiments must be repeated using different forum data to test if the same results are obtained or not, that is, if the EM clustering algorithm obtains again a high accuracy comparable with traditional classification algorithms. In the future, we hope to automate the process of evaluating student messages, because evaluating messages manually is a very difficult and time-consuming task for instructors. A data text mining algorithm could be used to automatically detect and classify types of messages and

evaluate them. Finally, we are working on improving our Moodle forum module. We hope to develop a network analysis tool to graphically depict the forum interaction (sociograms) and to identify further measures than the two currently used (centrality and prestige) to provide valuable information for predicting students' final marks.

Mahesh Singh, Anita Rani, Ritu Sharma, "An Optimised Approach For Student's Academic Performance By K-Means Clustering Algorithm Using Weka Interface" 2014. In this research paper we use K-Means clustering algorithm using weka interface. This study evaluates and predicts the student learning activities. We hope that the information generated after the implementation of k-means clustering technique using weka interface may be helpful for teachers as well as students. This research may improve the student academic performance and reduce the failing ratio by taking appropriate steps at the right time to improve and enhance the quality of education

Brijesh Kumar Baradwaj, Saurabh Pal, "Mining Educational Data to Analyze Students‟ Performance" 2011, the classification task is used on student database to predict the students division on the basis of previous database. As there are many approaches that are used for data classification, the decision tree method is used here. Information‟s like Attendance, Class test, Seminar and Assignment marks were collected from the student‟s previous database, to predict the performance at the end of the semester. This study will help to the students and the teachers to improve the division of the student. This study will also work to identify those students which needed special attention to reduce fail ration and taking appropriate action for the next semester examination.

## VI Problem Statement

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: 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.

## VII 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

- 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.
- We applied piecewise vector quantization method to transform dataset.
- 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.
- 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.
- We compared the distortion produced due to transformation of dataset by the distortion metric.

## VIII Experiment Analysis

Dataset Information

Water treatment 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.

**Table 1: Dataset Information**

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

## IX Steps for Dataset Procedure

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.

## X Results

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.

ISSN: 2581-3404 (Online)
International Journal of Innovative Research in Technology and Management (IJIRTM), Volume-3, Issue-6, 2019

**Table 2: Distortion value at different K and L value.**

| 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 1: Shows Distortion Vs Segment Size (L). Segment size varies from 2 to 19 and its corresponding distortion measured by distortion metric on various values 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.

ISSN: 2581-3404 (Online)
International Journal of Innovative Research in Technology and Management (IJIRTM), Volume-3, Issue-6, 2019

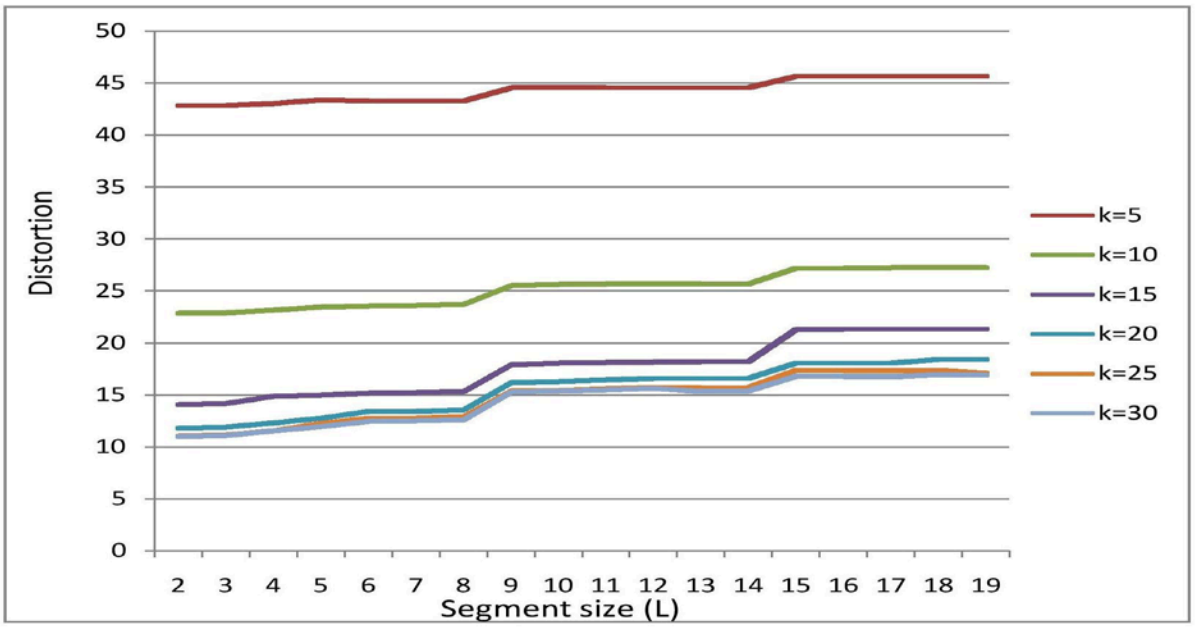


**Figure 1:** Line Graph of Distortion Vs Segment size (L) at different K value.

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 2 as a line graph and Figure 3 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.

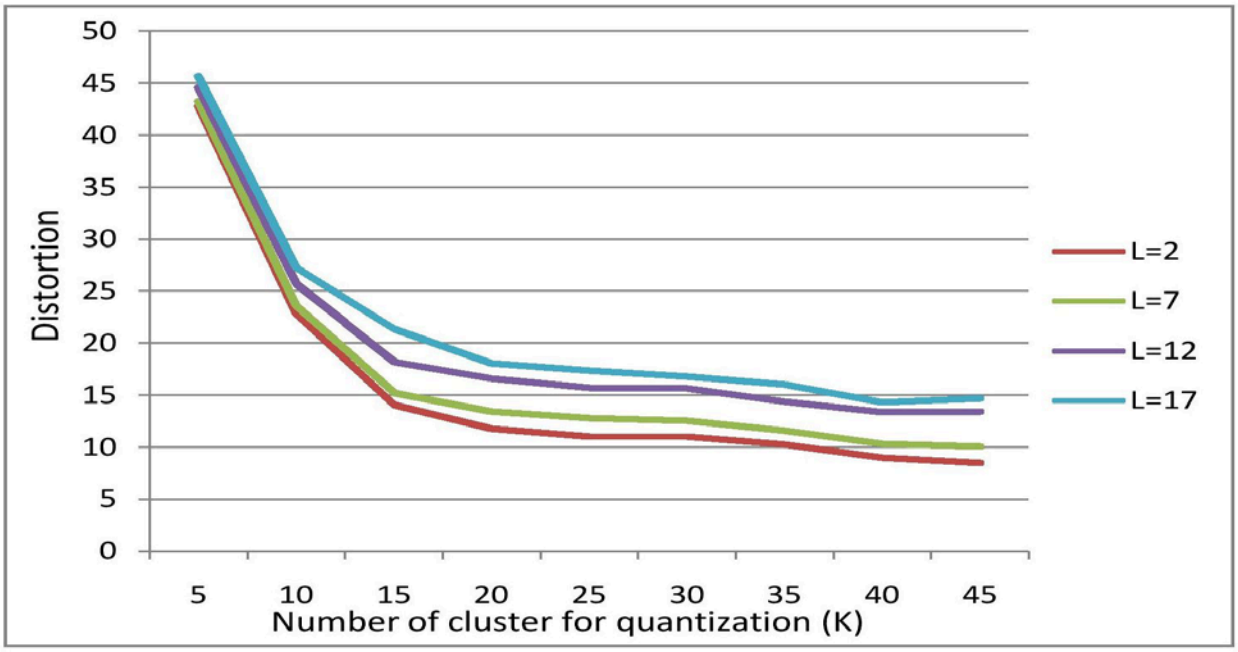


**Figure 2:** Line Graph of Distortion Vs Number of cluster for quantization (K) at different L.

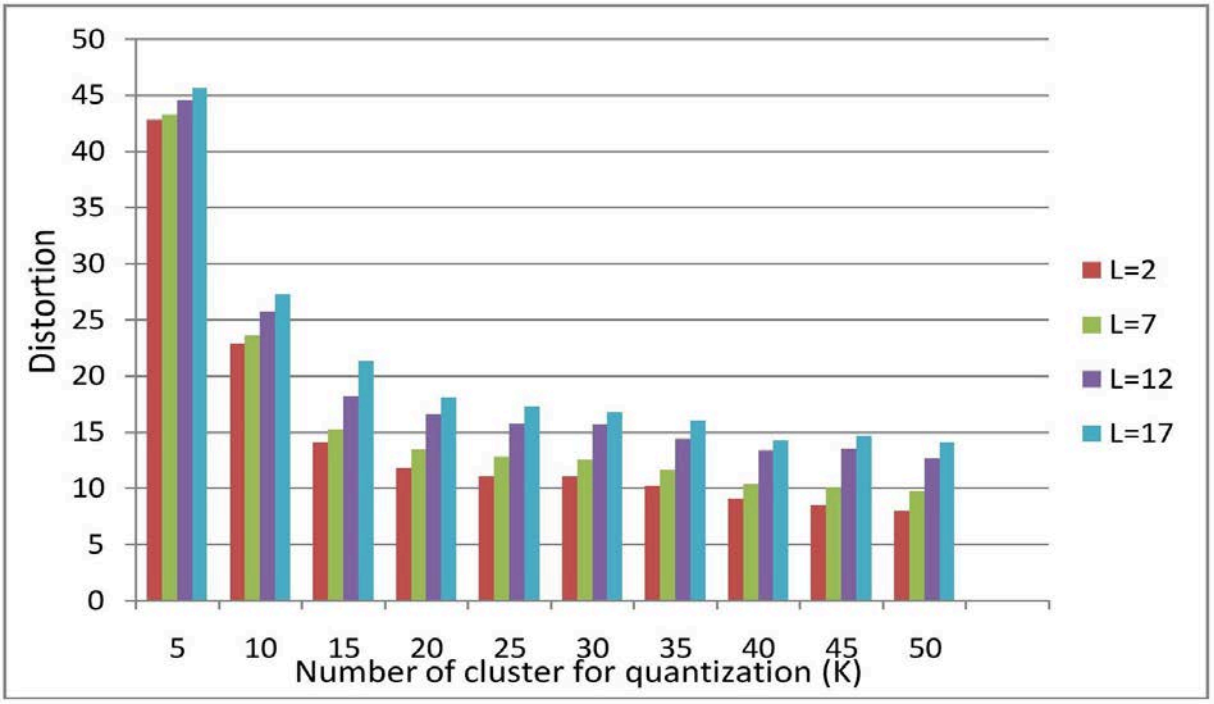


**Figure 3:** Column Graph of Distortion Vs Number of cluster for quantization (K) at different L.

## XI Conclusion

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 F-measure, Our experiment showed the variation of F-measure 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 F-measure and distortion and in turn privacy for water treatment dataset

**REFERENCES:-**

1. Agrawal R., Srikant R. Privacy preserving data mining. In: Proceeedings of the ACM SIGMOD Conference of Management of Data, pp. 439–450. ACM (2000).
2. Bramer Max, Principles of Data Mining, London, Springer, 2007.
3. 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.
4. Agarwal Charu C., Yu Philip S., Privacy Preserving Data Mining: Models and Algorithms, New York, Springer, 2008.
5. 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.
6. Wang Qiang , Megalooikonomou, Vasileios, A dimensionality reduction technique for efficient time series similarity analysis, Inf. Syst. 33, 1 (Mar.2008), 115- 132.
7. UCI Repository of machine learning databases, University of California, Irvine.