KONERU LAKSHMAIAH EDUCATION FOUNDATION
(Deemed to be university estd, u/s, 3 of the UGC Act, 1956)
(NAAC Accredited "A" Grade University)

PROJECT BASED LAB REPORT

On

K-Means clustering for telecommunication domain

Submitted in partial fulfilment of the

Requirements for the award of the Degree of

Bachelor of Technology

In

Computer science and Engineering

Under the esteemed guidance of

Dr. V.Bhavani

By

GG Krishna Vamsi (170030383)

(DST-FIST Sponsored Department)

K L EDUCATION FOUNDATION

Green Fields, Vaddeswaram, Guntur District-522 502

2019-2020

K L EDUCATION FOUNDATION

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

(DST-FIST Sponsored Department)

CERTIFICATE

This is to certify that this project based lab report entitled **" K-Means clustering for telecommunication domain** " is a bonafide work done by **GG Krishna Vamsi (170030383)** in the course **17CS3065 Big Data Analytics** in partial fulfilment of the requirements for the award of Degree in Bachelor of Technology in **COMPUTER SCIENCEAND ENGNEERING** during the Even Semester of Academic year 2019-2020.

Faculty in Charge **Head of the Department**

K L EDUCATION FOUNDATION

DEPT OF COMPUTER SCINCE AND ENGINEERING

(DST-FIST Sponsored Department)

DECLARATION

I hereby declare that this project based lab report entitled **"K-Means clustering for telecommunication domain"** has been prepared by **GG Krishna Vamsi (170030383)** in the course **17CS3065 & Big Data Analytics** in partial fulfilment of the requirement for the award of degree bachelor of technology in **COMPUTER SCIENCE AND ENGINEERING** during the Even Semester of the academic year 2019-2020. I also declare that this project-based lab report is of my own effort and it has not been submitted to any other university for the award of any degree.

Date:

Place:

ACKNOWLEDGEMENT

Our sincere thanks to **Dr. V.Bhavani** in the Lab for her outstanding support throughout the project for the successful completion of the work.

We express our gratitude to **Dr P.Vidyullatha** Course Co-ordinator for **17CS3065 Big Data Analytics** course in the Computer Science and Engineering Department for providing us with adequate planning and support and means by which we can complete this project-based Lab.

We express our gratitude to **Mr. V. HARIKIRAN,** Head of the Department for computer science and Engineering for providing us with adequate facilities, ways and means by which we can complete this project-based Lab.

We would like to place on record the deep sense of gratitude to the honourable Vice Chancellor, K L University for providing the necessary facilities to carry the project-based Lab.

Last but not the least, we thank all Teaching and Non-Teaching Staff of our department and especially our classmates and our friends for their support in the completion of our project-based Lab.

GG Krishna Vamsi (170030383)

TABLE OF CONTENTS

CHAPTER 1 - ABSTRACT

Cluster analysis is one of the major data analysis methods widely used for many practical applications in emerging areas of data mining. A good clustering method will produce high quality clusters with high intra-cluster similarity and low inter-cluster similarity. Clustering techniques are applied in different domains to predict future trends of available data and its uses for the real world. This research work is carried out on telecommunication domain using k-Means clustering algorithm. A state of art analysis of algorithm is implemented and performance is analyzed based on their clustering result quality by means of its execution time and other components. Telecommunication data is the source data for this analysis. The connection oriented broadband data is given as input to find the clustering quality of the algorithms. Distance between the server locations and their connection is considered for clustering. Execution time for each algorithm is analyzed and the results are compared with existing models. Results found in comparison study are satisfactory.

Keywords— k-Means Algorithm, Data Clustering, Time Complexity, Telecommunication Data

INTRODUCTION

Data Mining (DM) is a convenient way of extracting patterns, which represents knowledge implicitly stored in large data sets and focuses on issues relating to their feasibility, usefulness, effectiveness and scalability. Data mining approach and its technology is used to extract the unknown pattern from the large set of data for the business and real time applications. It can be viewed as an essential step in the process of knowledge discovery. Data are normally preprocessed through data cleaning, data integration, data selection, and data transformation and prepared for the mining task. Started as little more than a dry extension of DM techniques, DM is now bringing important contributions in crucial fields of investigations. Among the traditional sciences like astronomy, high energy physics, biology and medicine [1] have always provided a rich source of applications to data miners. An important field of application for data mining techniques is also the World Wide Web. The Web provides the ability to access one of the largest data repositories, which in most cases still remains to be analyzed and understood. Recently, DM techniques are also being applied to social sciences, homeland security and counter terrorism. A DM system is therefore composed of a software environment that provides all the functionalities to compose DM applications, and a hardware back-end onto which the DM applications are executed.

Data mining can be performed on various types of databases and information repositories, but the kind of patterns to be found are specified by various data mining functionalities like class/concept description, association, correlation analysis, classification, prediction, cluster analysis etc. Among these, Cluster analysis is one of the major data analysis method widely used for many practical applications in emerging areas [2]. Clustering is the process of finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups. The quality of a clustering result depends on both the similarity measure used by the method and its implementation and also by its ability to discover some or all of the hidden patterns. There is a number of clustering techniques that have been proposed over the years [3]. Different clustering approaches may yield different results. The Partitioning based algorithms are frequently used by many researchers for various applications in different domains. This research work compares two of the partitioning based clustering techniques namely k-Means and k-Medoids via its performance based on their execution time. The remainder of the paper is structured as follows. The

next section provides a comprehensive outline of related work via literature survey. Section 3 describes the basic approach and method of both algorithms. An experimental setup of the telecommunication data and the properties of the same data are discussed in Section 4. Section 5 explores the clustering process and obtained results of the algorithms. Finally, Section 6 contains the concluding remarks of the research work.

LITERATURE SURVEY

Nowadays, data clustering has attracted the attention of many researchers in different disciplines. It is an important and useful technique in data analysis. A large number of clustering algorithms have been put forward and investigated. The main advantage of clustering is that interesting patterns and structures can be found directly from very large data sets with little or none of the background knowledge. The cluster results are not subjective, but implementation dependent. Data Clustering has been addressed by many researchers and many clustering approaches have been explored and studied. A variety of data clustering algorithms are developed and applied for many application domains in the field of data mining. Clustering techniques have been applied to a wide variety of research problems. Hartigan provides an excellent summary of the many published studies reporting the results of cluster analyses [4]. For example, in the field of medicine, clustering diseases, cures for diseases, or symptoms of diseases can lead to very useful taxonomies. In the field of psychiatry, the correct diagnosis of clusters of symptoms such as paranoia, schizophrenia, etc. is essential for successful therapy. In archeology, researchers have attempted to establish taxonomies of stone tools, funeral objects, etc. by applying cluster analytic techniques. In general, whenever one needs to classify a "mountain" of information into manageable meaningful piles, cluster analysis is of great utility.

Bradley P.S and Fayyad describe refining Initial Points for k-Means Clustering in their paper [5]. They said that the practical approaches to clustering use an iterative procedure (e.g. k-Means, EM) which converges to one of numerous local minima. This paper presents a procedure for computing a refined starting condition from a given initial one that is based on an efficient technique for estimating the modes of a distribution. Recently, Bhukya et al., deals with the performance evaluation of partition based clustering algorithms in grid environment using design of experiments [6]. In their work, they focus mainly on theanalysis of k-Means and k-Medoids algorithms. One of the disadvantages of using these algorithms is its unsuitability for larger data sets. To solve this problem Grid environment has been selected. The main objective of the work is to implement the partition based clustering algorithms in the Grid environment on Grid Gain middleware and analyze their performance for large datasets with Design of Experiment (DOE) framework. Finally, they conclude that the k-Means clustering algorithm is faster than k-Medoids when tested with large data sets and the results are found to be satisfactory. A review of the most common partition algorithms in cluster analysis: a comparative study is discussed in

a research work by Susana et al., in [7]. In this work, a simulation study was performed to compare the results obtained from the implementation of the algorithms k-means, k-medians, PAM and CLARA when continuous multivariate information is available. Additionally, a study of simulation is presented to compare partition algorithms qualitative information, comparing the efficiency of the PAM and k-modes algorithms. The efficiency of the algorithms is compared using the Adjusted Rand Index and the correct classification rate. Finally, the algorithms are applied to real databases with predefined classes.

PROPOSED SYSTEM

An Enhanced k-means algorithm to improve the Efficiency Using Normal Distribution Data Points is discussed by Napoleon and Ganga Lakshmi in their research work [8]. This paper proposes a method for making the k-means algorithm more effective and efficient; so as to get better clustering with reduced complexity. In this research, the most representative algorithms k-Means and the Enhanced k-Means were examined and analyzed based on their basic approach. They found that the elapsed time taken by proposed enhanced k-means is less than k-means algorithm. A work carried out by Benderskaya et al., titled as "Self-organized Clustering and Classification: A Unified Approach via Distributed Chaotic Computing" [9] describes a unified approach to solve clustering and classification problems by means of oscillatory neural networks with chaotic dynamics. The advantages of distributed clusters formation in comparison to centers of clusters estimation are demonstrated. New approach to clustering on-the-fly is proposed.

CHAPTER 2: REQUIREMENTS & ANALYSIS

Platform Requirements

Hardware/Software	**Hardware / Software element**	**Specification /version**
Hardware	Processor	Intel i5
	RAM	8GB
	Hard Disk	1TB
Software	OS	Windows 10
	R Studio	cloud

Modules Description

There are few common steps involved in implementing the algorithms as

- **Data Collection**: The data we have Gathered from the telecommunication center to analyze the telecom data.
- **Data Preprocessing**: It involves removing the outliers of the data, Eliminating the unnecessary data
- **Data Transformation**: It is the process of converting data from one format or structure into another format or structure.
- **Applying Algorithms**: It involves the application of the data mining algorithms to the given dataset.
- **Predicting the data**: It is the practice of extracting information from existing data sets in order to determine patterns and predict future outcomes and trends.

CHAPTER 3: DESIGN & IMPLEMENTATION

ALGORITHMS

The k-Means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori [14]. This algorithm aims at minimizing an objective function, in this case a squared error function. The objective function is as follows :

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

. The algorithm is composed of the following steps:

Step 1: Place k points into the space represented by the objects that are being clustered. These points represent initial group centroids.

Step 2: Assign each object to the group that has the closest centroid.

Step 3: When all objects have been assigned, recalculate the positions of the k centroids.

Step 4: Repeat Steps 2 and 3 until the centroids no longer move.

This produces a separation of the objects into groups from which the metric to be minimized can be calculated. Although it can be proved that the procedure will always terminate, the k-Means algorithm does not necessarily find the most optimal configuration, corresponding to the global objective function minimum [15]. This k-Means is a simple algorithm that has been adapted to many problem domains [16].

PSEUDOCODE

```
library(tidyverse)
library(MASS)
library(car)
library(e1071)
library(caret)
library(cowplot)
library(caTools)
library(pROC)
library(ggcorrplot)
telco <- read.csv("../input/WA_Fn-UseC_-Telco-Customer-Churn.csv")
glimpse(telco)
telco <- telco[complete.cases(telco),]
telco$SeniorCitizen <- as.factor(ifelse(telco$SeniorCitizen==1, 'YES', 'NO'))
options(repr.plot.width =6, repr.plot.height = 2)
ggplot(telco, aes(y= tenure, x = "", fill = Churn)) +
geom_boxplot()+
theme_bw()+
xlab(" ")
ggplot(telco, aes(y= MonthlyCharges, x = "", fill = Churn)) +
geom_boxplot()+
theme_bw()+
xlab(" ")
ggplot(telco, aes(y= TotalCharges, x = "", fill = Churn)) +
geom_boxplot()+
theme_bw()+
xlab(" ")
options(repr.plot.width =4, repr.plot.height = 4)
boxplot(telco$tenure)$out
boxplot(telco$MonthlyCharges)$out
boxplot(telco$TotalCharges)$out
telco <- data.frame(lapply(telco, function(x) {
                gsub("No internet service", "No", x)}))

telco <- data.frame(lapply(telco, function(x) {
                gsub("No phone service", "No", x)}))
num_columns <- c("tenure", "MonthlyCharges", "TotalCharges")
telco[num_columns] <- sapply(telco[num_columns], as.numeric)

telco_int <- telco[,c("tenure", "MonthlyCharges", "TotalCharges")]
telco_int <- data.frame(scale(telco_int))

telco <- mutate(telco, tenure_bin = tenure)

telco$tenure_bin[telco$tenure_bin >=0 & telco$tenure_bin <= 12] <- '0-1 year'
telco$tenure_bin[telco$tenure_bin > 12 & telco$tenure_bin <= 24] <- '1-2 years'
```

```
telco$tenure_bin[telco$tenure_bin > 24 & telco$tenure_bin <= 36] <- '2-3 years'
telco$tenure_bin[telco$tenure_bin > 36 & telco$tenure_bin <= 48] <- '3-4 years'
telco$tenure_bin[telco$tenure_bin > 48 & telco$tenure_bin <= 60] <- '4-5 years'
telco$tenure_bin[telco$tenure_bin > 60 & telco$tenure_bin <= 72] <- '5-6 years'

telco$tenure_bin <- as.factor(telco$tenure_bin)
options(repr.plot.width =6, repr.plot.height = 3)
ggplot(telco, aes(tenure_bin, fill = tenure_bin)) + geom_bar()+ theme1
telco_cat <- telco[,-c(1,6,19,20)]

#Creating Dummy Variables
dummy<- data.frame(sapply(telco_cat,function(x) data.frame(model.matrix(~x-1,data =telco_cat))[,-
1]))

head(dummy)
#Combining the data
telco_final <- cbind(telco_int,dummy)
head(telco_final)
#Splitting the data
set.seed(123)
indices = sample.split(telco_final$Churn, SplitRatio = 0.7)
train = telco_final[indices,]
validation = telco_final[!(indices),]
model_1 = kmeans(telco_final,centers=2, iter.max = 25, nstart=100)
summary(model_1)
cutoff_churn <- factor(ifelse(pred >=0.32, "Yes", "No"))
conf_final <- confusionMatrix(cutoff_churn, actual_churn, positive = "Yes")
accuracy <- conf_final$overall[1]
sensitivity <- conf_final$byClass[1]
specificity <- conf_final$byClass[2]
accuracy
sensitivity
specificity
```

CHAPTER 4: SCREENSHOTS & RESULTS

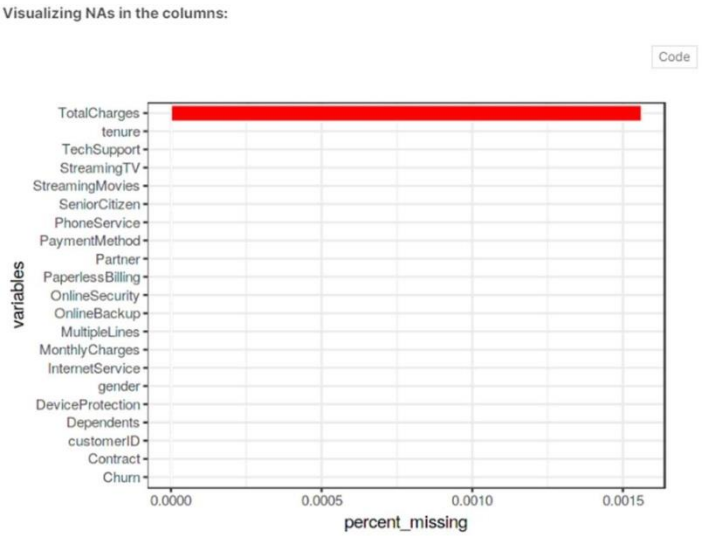

Fig 1.0 Columns in the Dataset

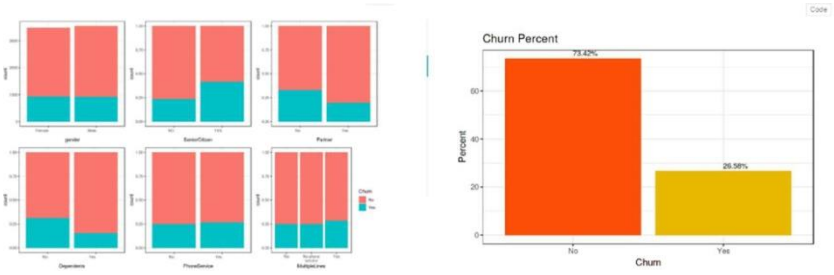

Fig 1.1 Columns Visualization

Fig 1.2 Box plot Visualisation

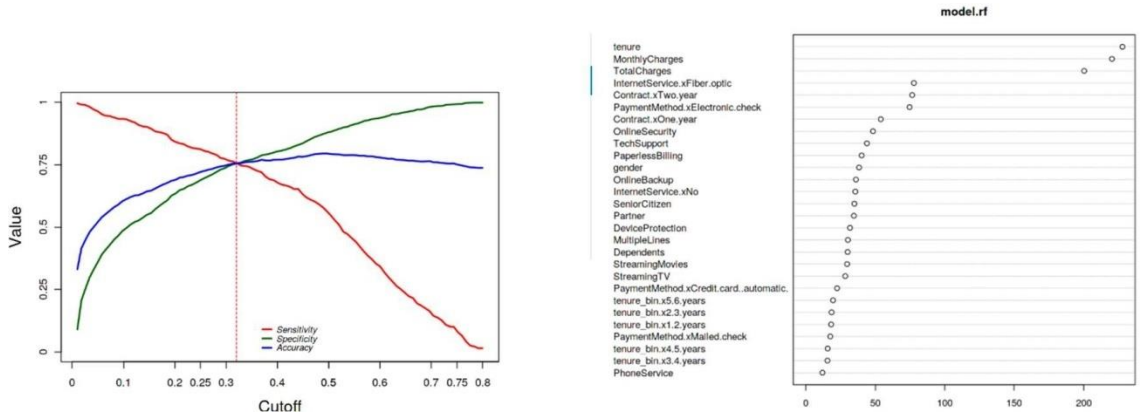

Fig 1.3 Perfomance Metrics

Output Visualizations

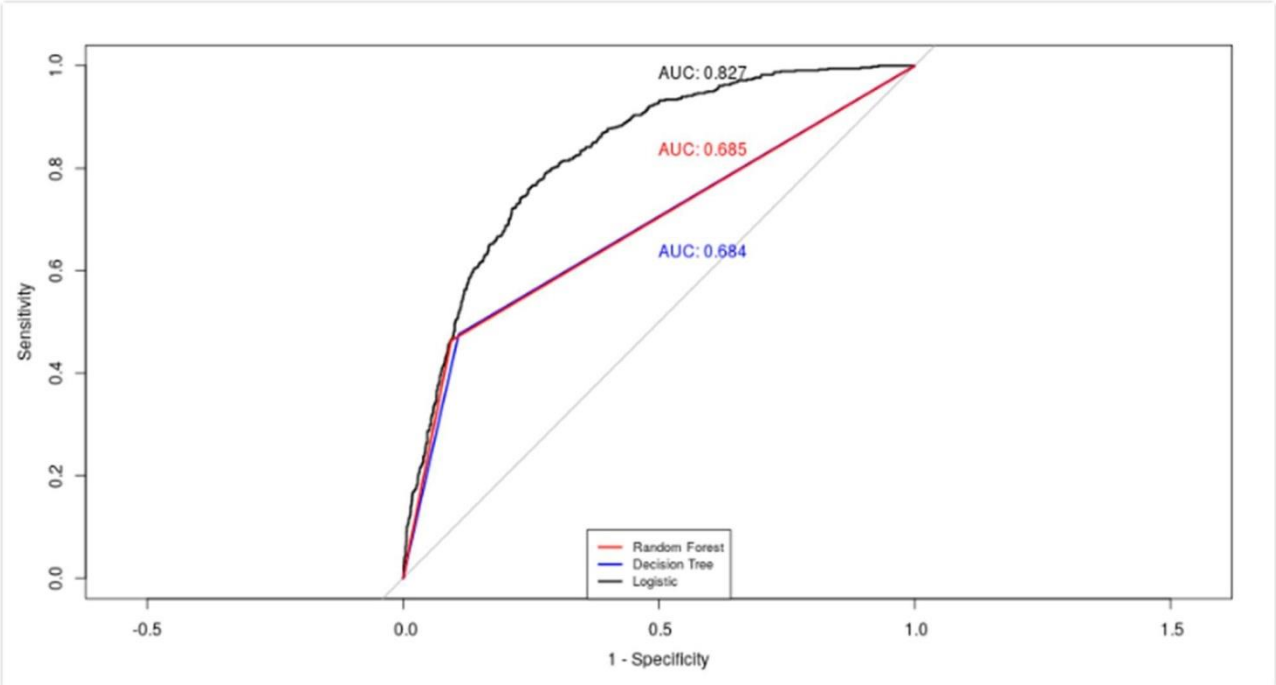

Fig 1.4 Roc-auc curves

CHAPTER 5: CONCLUSION

Cluster analysis is still an active field of development. Many cluster analysis techniques do not have a strong formal basis. In summary, clustering is an interesting, useful, and challenging problem. It has great potential in applications like object recognition, image segmentation, and information filtering and retrieval. From the experimental approach, by several executions of the program, proposed algorithms in this research work, following results were obtained. The advantage of the k-Means algorithm is its favorable execution time. Its drawback is that the user has to know in advance how many clusters are searched for.

CHAPTER 6: REFERENCES

[1] Han, J. and Kamber, M. (2006) Data Mining: Concepts and Techniques. 2nd Edition, Morgan Kaufmann Publishers, New Delhi.

[2] Jain, A.K., Murty, M.N. and Flynn, P.J. (1999) Data Clustering: A Review. ACM Computing Surveys, 31. https://doi.org/10.1145/331499.331504

[3] Berkhin, P. (2002) Survey of Clustering Data Mining Techniques. Technical Report, Accrue Software, Inc.

[4] Hartigan, J.A. (1975) Clustering Algorithms. Wiley Publishers.

[5] Bradley. P.S., Fayyad, U.M. and Reina, C.A. (1998) Scaling Clustering Algorithms to Large Databases. Proceedings of the 4th International Conference on Knowledge Discovery & Data Mining, AAAI Press, Menlo Park, CA, 9-15 .

[6] Bhukya, D.P., Ramachandram, S. and Reeta Sony, A.L. (2010) Performance Evaluation of Partition Based Clustering Algorithms in Grid Environment Using Design of Experiments. International Journal of Reviews in Computing, 4, 46-53.

[7] Leiva-Valdebenito, S.A. and Torres-Aviles, F.J. (2010) A Review of the Most Common Partition Algorithms in Cluster Analysis: A Comparative Study. Colombian Journal of Statistics, 33, 321-339.

[8] Napoleon, D. and Ganga Lakshmi, P. (2010) An Enhanced k-Means Algorithm to Improve the Efficiency Using Normal Distribution Data Points. Int. Journal on Computer Science and Engineering, 2, 2409-2413 Segmentation. Journal of Computer Science, 7, 657-663.