Asian Journal of Computer Science and Technology
ISSN: 2249-0701 Vol.8 No.S2, 2019, pp. 7-8

Airline Traffic Analysis Using Clustering Method in R Language

S.Thanganila[1], T. Keerthana[2] and P.Tamilzhchelvi[3]
[1&2]PG Student, [3]Associate Professor
[1,2&3]Department of Computer Science, Ayya Nadar Janaki Ammal College, Sivakasi, Tamil Nadu, India
E-Mail: thanganila54@gmail.com,keerthana220296@gamail.com,tamilindhu@rediffmail.com

***Abstract* -In this paper, we analyze the airlines traffic in R using k-means clustering algorithm to avoid traffic in airlines. Clustering is the process of grouping data, where grouping is recognized by discovering similarities between data based on their features. K-Means algorithm is applied to get results and to predict before any problem arise in airline traffic. R takes too much time to load huge amount of data and sometimes does not support to upload huge volume of datasets. So, to avoid this space and time complexity, we have used Google Cloud.**
***Keywords:* Cluster Analysis, Airline Data, K-Means Algorithm, Google Cloud**

I. INTRODUCTION

The R programming language provides wide selection of graphical & statistical techniques such as linear modeling, nonlinear modeling, and classification; clustering etc. Clustering is the part of the unsupervised learning which groups particular set of objects based on their characteristics and aggregating them according to their similarities[2]. We use Google cloud in order to work with huge volume of dataset. Google cloud allows to deploy our workload on virtual machines.

A. Overview of the Paper: This paper uses a data set that contains a lot of information about the traffic flow of Europe city. This data when mined over location can provide information about the major attractions of the city.

The data when monitored over time can help us to identify rush hours, holiday season, impact of weather, etc. This knowledge can be applied for better planning and traffic management.

B. Aim and Objective of Paper: To prevent and avoid airline traffic using k-means clustering algorithm

II. MATERIALS AND METHODS

There are a number of clustering algorithms that have been proposed by several researchers in the field of clustering applications [5]. Such algorithms create high impact in their clustering result quality. This research work deals with partition based clustering algorithm namely k-means.

A.The k-Means Algorithm

K-means is one of the simplest unsupervised learning algorithms that solve clustering problem [6]. The procedure follows a simple and easy way to classify a given dataset through a certain number of clusters (assume k clusters) fixed apriori. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as bary center of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more. Finally, this algorithm aims at minimizing an objective function knows as squared error function given by:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\|x_i - v_j\| \right)^2$$

Where,

'$||x_i - v_j||$' is the Euclidean distance between x_i and v_j.

'c_i' is the number of data points in i^{th} cluster.

'c' is the number of cluster centers.

III. IMPLEMENTATION

K-means algorithm is used here to cluster the Europe Airline Dataset and this huge dataset is stored in Hadoop platform and retrieved through R Hadoop client Package in R. We apply k-means function in R for clustering the traffic flow and the result is displayed using GGMAP based on the latitude and longitude. Thus, we have shown a way to prevent airline traffic and delay of flights.

A. Properties of the Dataset - As of January 2017, the Open Flights Airports Database contain traffic flow of more than 10,000 airports, train stations and ferry terminals spanning the globe [3]. Each entry contains the following information.

Dataset size: 2MB

Airport ID: Unique Open Flights identifier for this airport.

Name: Name of airport. May or may not contain the City name.
City: Main city served by airport. May be spelled differently from name.
Country: Country or territory where airport is located.
Latitude : Decimal degrees, usually to six significant digits. Negative is South, positive is North.
Longitude: Decimal degrees, usually to six significant digits. Negative is West, positive is East.
Latitude : In feet.
Time Zone: Hours offset from UTC. Fractional hours are expressed as decimals.

B. Platform Preparation

Step1: create an account in google cloud platform using Gmail id.
Step2: And create a Hadoop cluster using Dataproc instance.
Step3: In dataproc instance chooses any one Hadoop framework and configures the instance.
Step 4: we store the airline dataset using dataprocssh.
Step5: [9] Install R studio in Hadoop cluster using following commands.

$ sudo apt-get install gdebi-core
$ sudo apt-get install libapparmor1
$ wget http: //download2.rstudio.org/rstudio-server-0.98.1102-amd64.deb
$ sudogdebi rstudio-server-0.98.1102-amd64.deb
$ sudorstudio-server verify-installation

Step 6: Login with username and password.
Step 7: In Rstudio install the required packages.

Install. Packages(“”)

The above function is used to install the packages.
Step8: Retrieve the data from Hadoop using [8] rhadoopclient package.
Step 9: We use kmeans () to cluster the output based upon their longitude and latitude value.
Step 10: To plot the output using ggmap map package.

IV. RESULT

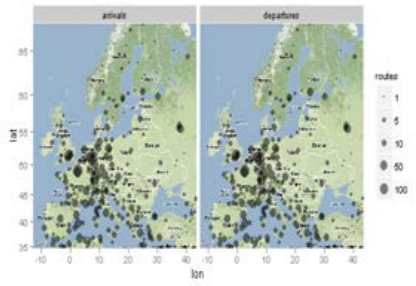

Fig.1 Cluster Output

V.CONCLUSION

In this paper we analyze the all flight plans for flight into, out of and around Europe are clustered. The above result is used to identify the traffic location easily and give the solution to the poilets and chance the routes in rush hour.

REFERENCES

[1] Akinori Harada, Tooru Ezaki, Tomoaki Wakayama, and Koichi Oka, “Air Traffic Efficiency Analysis of Airliner Scheduled Flights Using Collaborative Actions for Renovation of Air Traffic Systems”, *Open Data.2017*.
[2] (2017) Stat.ethz.ch website. [Online] Available at: https://stat.ethz.ch/R-manual/R-devel/library/stats/html/kmeans.html
[3] (2012) data camp website [Online] Available at: https://www.data camp.com/community/tutorials/k-means-clustering-r.
[4] Airline dataset, [Online] Available at: https://openflights.org/data. html.
[5] (2015) rblogger [Online] Available at: https://www.r-bloggerscom/k-means-clustering-in-r/,
[6] (2016) datanovia [Online] Available at: https://www.datanovia.Com/en/ lessons/k-means-clustering-in-r-algorith-and-practical-examples/,
[7] Sites. google [Online] Available at: https://sites.google.com/site/data clusteringalgorithms/k-means-clustering-algorithm.
[8] (2010)CRAN, [Online] Available at: https://cran.r-project.org/web/packages/rHadoopClient/index.html.
[9] (2012). Medium, [Online] Available at: https://medium.com/@GalarnykMichael/install-r-and-rstudio-on-ubuntu-12-04-14-04-16-04-b6b3107f7779.