Cluster Analysis of Trees for Fauna

Subham Chauhan, Swarnim Gupta, Nipun Tank

Abstract: ***Nowadays, many problems are arising due to large scale deforestation and soil erosion. Instead of leaving the land uncultivated, it can be used in a beneficial way by growing beneficial plants. Different species of plants require different climatic conditions for optimum growth. So, clustering of plants can be beneficial such that plants will get an adequate amount of nutrients. By clustering the plants, the appropriate land can be found in which surplus amount can be grown easily. Different plants need different types of nutrients from the soil and also, the groundwater level of all areas isn't the same. Hence, there is a need to identify the right area of land for cultivation. This will not only be advantageous for humans and animals but also, it will help improve the environment of the area. Afforestation will help improve the water cycle, reduce soil erosion and other such issues. The report will consist of trees grouped in clusters of different species. This could be done by data mining, and the clustering algorithm is a particular data mining concept. Cluster analysis is a method in which clusters are formed based on common characteristics i.e. elements in the same cluster are analogous to each other than those in other clusters. R language could be used for clustering trees and plants of a particular area. It is free software for statistical computing and graphics.***

Index Terms: ***Cluster Analysis, Data mining, Land Utilization, R tool***

I. INTRODUCTION

Topically, collection and analysis of copse-related data are pertinent to ICAR (Indian Council of Agricultural Research). The use of a coherent method to classify these data based on the climatic condition and location of uncultivated land, types of trees requiring similar natural resources, and prediction of future utilization of the land are the most substantive aspects that have to be addressed.

DATA MINING

Data mining is the way toward grouping information by discovering designs in vast data collections. By arranging information into smaller gatherings with comparable attributes, it turns out to be considerably simpler to deal with the data in a productive way. The expression "Data mining" is in truth an off base name, in light of the fact that the objective is the extraction of examples and learning from a lot of information, not simply the extraction of information. It likewise is a prominent term and is every now and again connected to any type of expansive scale information or data handling (gathering, extraction, warehousing, examination, and measurements) and in addition any use of PC choice emotionally supportive network, including computerized reasoning (e.g., machine learning) and business knowledge. The learning disclosure in databases (KDD) process is normally characterized with the stages:

1. Selection
2. Pre-preparing
3. Transformation
4. Data mining
5. Interpretation/assessment.

K-NEAREST NEIGHBORS ALGORITHM

After pre-preparing information, we can utilize different bunching calculations. K-implies bunching intends to segment n perceptions into k groups in which every perception has a place with the group with the closest mean, filling in as a model of the group. This outcomes in a dividing of the information space into Voronoi cells. This calculation fills in as takes after:

1.Register the Euclidean separation from the inquiry case to the marked cases.

2. Request the named cases by expanding separation.

II. LITERATURE SURVEY

Rasoul Kiani et al.[1] propose a new framework for clustering and predicting crimes based on real data. In this paper, GA was used to improve outlier detection in the pre-processing phase, and the fitness function was defined based on accuracy and classification error parameters. Hence, the fitness function was optimized.

Radha Mothukuri et al.[2] propose a methodology to provide security for crime data during outsourcing. Clustering and classification is made on the crime information. While classifying, watermark content is added for the purpose of defence to verify classification data.

Arit Thammano[3] describes the most popular clustering algorithm because of its efficiency and superior performance. However, the performance of K-means algorithm depends heavily on the selection of initial centroids. This paper proposes an augmentation to the first K-implies calculation empowering it to take care of order issues.

Ying Zhao, George Karypis[4] depicts that quick and top notch report bunching calculations assume a critical part in giving natural route and perusing components by sorting out a lot of data into few important groups. This paper centers around archive bunching calculations that assemble such various leveled arrangements and (I) introduces a complete investigation of segment and agglomerative calculations that utilization diverse rule capacities and blending plans. (ii) exhibits another class of bunching calculations called obliged agglomerative calculations.

A. Malathi et al.[5] center around advancement of a crime analysis utilizing distinctive data mining systems to help anticipate the violations examples and quick up the

Revised Manuscript Received on May 06, 2019

Subham Chauhan, CSE, SRM Institute of Science and Technology, Chennai, India.

Swarnim Gupta, CSE, SRM Institute of Science and Technology, Chennai, India.

Nipun Tank, CSE, SRM Institute of Science and Technology, Chennai, India.

way toward settling wrongdoing.

There are four stages, to be specific, information cleaning, grouping, order and exception location. These systems joined with cutting edge PCs can be utilized to survey to a great degree substantial datasets, in this manner sparing time.

A. Malathi et al.[6] utilized a grouping/characterize based model to envision crime analysis patterns. The information mining strategies are utilized to break down the city crime analysis data from Police Department. The consequences of this data mining could possibly be utilized to diminish and even avoid crime for the expected years.

Swadi Al-Janabi[7] presents a proposed structure for the crime and criminal information examination and identification utilizing choice tree calculations for information grouping and straightforward K-implies calculation for information bunching. The paper tends to help pros in finding examples and patterns, making gauges, discovering connections and conceivable clarifications.

Aravindan Mahendiran et al.[8] apply bunch of apparatuses on crime data indexes to dig for data that is avoided human recognition. With the assistance of best in class representation systems, they exhibit the examples found through their calculations in a slick and natural way that empowers law requirement divisions to channelize their assets as needs be.

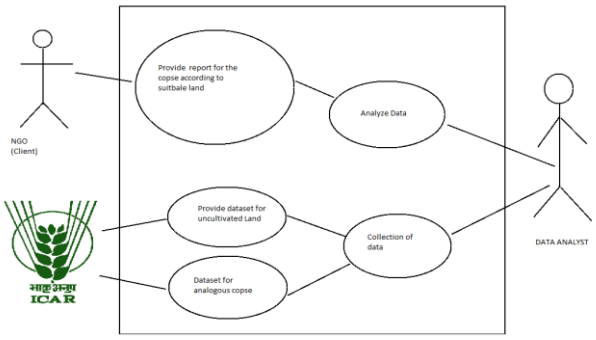
Sutapat Thiprungsri et al.[9] look at the use of cluster analysis in the bookkeeping space, specifically, its application to error location in the field of review. Computerizing extortion sifting can be of incredible incentive to ceaseless reviews. The goal of their investigation is to look at the utilization of group examination as an option and creative irregularity discovery method in the wire exchange framework.

Chun-Nan Hsu, Han-Shen Huang, Bo-Hou Yang[10] depict the Expectation-Maximization(EM) calculation as a standout amongst the most prominent calculations for data mining from fragmented data. Nonetheless, when connected to substantial informational indexes with an extensive extent of missing information, the EM calculation may unite gradually.

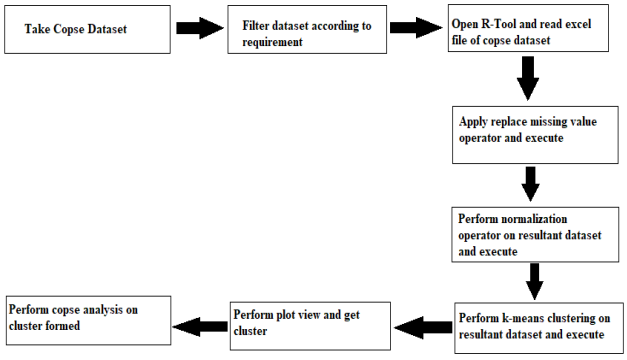
III. METHODS

Data pre-handling is a data mining procedure that includes changing crude information into a reasonable configuration. Frequently the data is unstructured, conflicting, has missing qualities, and need in certain conduct or patterns that gives numerous mistakes. In this manner, it should be cleaned, coordinated, changed, and thus lessened. Cleaning fills in the missing qualities and expels raucous. Change utilizes standardization and totals the information and Reduction helps in diminishing the volume of information keeping comparative systematic outcomes.

In the existing system, the data was collected manually after surveying the land. From the graph data, we can synthesize datasets about the uncultivated lands and other about the climatic conditions that land provides. Using these, a list of tress which can acclimate to the corresponding land can be generated. It will be helpful for utilizing the barren lands by growing the trees and thus, preventing soil erosion and degradation of land. After surveying the particular area, we could cull it according to the apt requisite.



For implementation purpose, we created sample data and then, applied clustering techniques to provide output graph.



IV. ALGORITHM

First, we will install all the libraries -

install.packages()

Then, read data, step by step, of various months with the help of -

read.csv

Let's bind all the data files into one. For this, you can use the bind_rows() function under the dplyr library in R.

We use aggr() function to check the missing values.

Then, with the help of kmeans() function we clustered our given data.

Next, we made gg-plot based on the usage of trees.

Code –

```
setwd("G:\\csv")
year2016<- read.csv("year2016.csv")
year2017<- read.csv("year2017.csv")
year2018<- read.csv("year2018.csv")
install.packages("dplyr")
library(dplyr)
mydata <- bind_rows(year2016,year2017,year2018)
summary(mydata)
install.packages("VIM")
library(VIM)
aggr(mydata)
set.seed(20)
clusters <- kmeans(mydata[,2:3],3)
mydata$Usage <- as.factor(clusters$cluster)
str(clusters)
library(ggplot2)
```

qplot(data=mydata, x= Area.Covered, y=Growth.Rate, colour=(as.factor(Usage)), size=(10))
mydata

Output –

```
> summary(mydata)
       Tree.Name   Area.Covered    Growth.Rate
 Almond      : 1   Min.   :0.340   Min.   : 7.53
 Aloe Vera   : 1   1st Qu.:1.280   1st Qu.:22.45
 Bahrain     : 1   Median :2.300   Median :35.80
 Banana      : 1   Mean   :3.059   Mean   :43.39
 Banyan      : 1   3rd Qu.:4.300   3rd Qu.:72.54
 Burkina Faso: 1   Max.   :9.000   Max.   :90.54
 (Other)     :23
>
```

This is the summary of the 'mydata', which we got after binding all the rows of the three files for
different consecutive years.

```
List of 9
 $ cluster     : int [1:29] 1 2 2 3 3 3 2 3 2 3 ...
 $ centers     : num [1:3, 1:2] 2.9 3.42 2.9 41.7 78.61 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:3] "1" "2" "3"
  .. ..$ : chr [1:2] "Area.Covered" "Growth.Rate"
 $ totss       : num 20024
 $ withinss    : num [1:3] 274 529 867
 $ tot.withinss: num 1669
 $ betweenss   : num 18354
 $ size        : int [1:3] 7 9 13
 $ iter        : int 2
 $ ifault      : int 0
 - attr(*, "class")= chr "kmeans"
>
```

This is the summary of the 3 clusters, which we formed with the help of the kmeans() function.

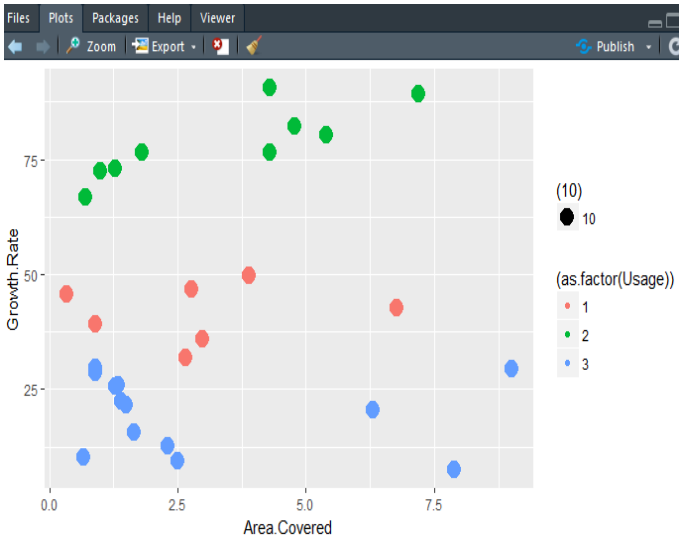
Untitled1* | mydata | year2016 | year2017

Filter

	Tree.Name	Area.Covered	Growth.Rate	Usage
1	Mint	0.90	39.22	1
2	Neem	1.80	76.56	2
3	Olive	5.40	80.23	2
4	Pecan	1.50	21.50	3
5	Coconut	2.30	12.56	3
6	Aloe Vera	0.90	28.50	3
7	Sandalwood	4.30	76.54	2
8	Sagun	2.50	9.34	3
9	Banana	0.99	72.54	2
10	Teak	1.34	25.87	3
11	Champak	0.70	66.79	2
12	Banyan	6.30	20.56	3
13	Willow	7.20	89.24	2
14	Tecomella	1.40	22.45	3
15	Burkina Faso	7.89	7.53	3
16	Mesua Ferrera	2.78	46.87	1
17	Oak	9.00	29.45	3
18	Bahrain	4.30	90.54	2
19	Almond	6.77	42.60	1
20	Seasem	0.90	29.60	3
21	Drumstick	2.66	31.86	1
22	Gulmohar	4.77	82.10	2
23	Coral	2.98	35.80	1
24	Cork	1.66	15.55	3

Showing 1 to 25 of 29 entries

This is data representation of the clustered, which we got after performing K-means algorithm.
We created an additional column 'Usage' which has 3 different data values such as:-
Number 1 states that it has low usage
Number 2 states that it has average usage
Number 3 states that it has high usage



Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

RESULT-

On the basis of clustered data, we plotted the above graph. From the graph, one can easily analyse the data. In the graph, we plotted X-coordinate as Area Covered and Y-coordinate as Growth rate, from which the user can get to know Usage of the tree by looking at the plot. This helps the user to know which tree is more beneficial to be grown in the remaining uncultivated land.

V. CONCLUSION AND FUTURE ENHANCEMENT

This paper proposes a new framework for predicting clusters based on the types of trees present in a particular area. By knowing the type and number of different species of trees growing in a particular area, and how advantageous each tree is to fauna, we can use the uncultivated land in that area to grow the most beneficial trees. This will be further useful for the climate as well. The only disadvantage of this framework is that since it is a new concept, no proper dataset is currently available which can be used for implementation of this logic in real world. Hence, we need a proper survey of areas according to the fields used in this project. The project can be extended in future by gathering data about the soil of the area and thus, basing the technique in agricultural field which can help farmers to increase their crop production.

REFERENCES

1. Rasoul Kiani, Siamak Mahdavi and Amin Keshavarzi, Islamic Azad University, Marvdasht, Iran – "Analysis and Prediction of Crimes by Clustering and Classification", 2015
2. Radha Mothukuri, Dr. Bobba Basaveswara Rao, Acharya Nagarjuna University, Andhra Pradesh – "Cluster Ananalysis of Cyber Crime Data using R", 2018
3. J. Han and M. Kamber – "Data Mining: Concepts and Techniques", second ed. Morgan Kaufmann, 2006.
4. C.C. Aggarwal and P.S. Yu - "Finding Generalized Projected Clusters in High Dimensional Spaces", Proc. 26th ACM SIGMOND Int'l Conf. Management of Data, pp. 70-81, 2000.
5. A.Malathi ,Dr. S. Santhosh Baboo, D.G. Vaishnav College, Chennai – "Algorithmic Crime Prediction Model Based on the Analysis of Crime Clusters", 2011
6. Malathi. A ,Dr. S. Santhosh Baboo and Anbarasi . A – "An Intelligent Analysis of a City Crime Data Using Data Mining", 2011
7. Kadhim B. Swadi Al-Janabi, Department of Computer Science, Faculty of Mathematics and Computer Science, University of Kufa/Iraq – "A Proposed Framework for Analyzing Crime Data Set using Decision Tree and Simple K-means Mining Algorithms", 2011
8. Aravindan Mahendiran, Michael Shuffett, Sathappan Muthiah, Rimy Malla, Gaoqiang Zhang – "Forecasting Crime Incidents using Cluster Analysis and Bayesian Belief Networks", 2011
9. Sutapat Thirprungsri, Miklos A. Vasarhelyi, Rutgers University, USA – "Cluster Analysis for Anomaly Detection in Accounting Data : An Audit Approach", 2011
10. K.Kailing, H.p. Kriegel, P. Kroger and S. Wanka – "Ranking Interesting Subspaces for Clustering High Dimensional Data", Proc. Seventh European Conf. Principles and Practice of Knowledge Discovery in Databases (PKDD), pp. 241-252, 2003