Implementation of Clustering techniques using - cluster, factoextra, magrittr packages for USArrests dataset.

Expt No: 8 May 15,2019
Author: Subalakshmi Shanthosi S (186001008)

Aim

Implementation of different clustering techniques like - Partitioning and Hierarchical using R packages- cluster, factoextra, magrittr.

Description

1. What is Clustering?
    - Clustering is the classification of data objects into similarity groups (clusters) according to a defined distance measure.
    - It is used in many fields, such as machine learning, data mining, pattern recognition, image analysis, genomics, systems biology, etc.
    - Machine learning typically regards data clustering as a form of unsupervised learning.
2. Why Clustering and Data Mining in R?
    - Efficient data structures and functions for clustering.
    - Reproducible and programmable.
    - Comprehensive set of clustering and machine learning libraries.
    - Integration with many other data analysis tools.
3. Advantages and Disadvantages of clustering methodologies:

Algorithm	Advantages	Disadvantages
Partitioning Clustering Algorithm	1. Relatively scalable and simple. 2.Suitable for datasets with compact spherical clusters that are well-separated	1. Severe effectiveness degradation in high dimensional spaces 2. Poor cluster descriptors 3.Reliance on the user to specify the number of clusters in advance 4. High sensitivity to initialization phase, noise and outliers 5. Inability to deal with non-convex clusters of varying size and density.
Hierarchical Clustering Algorithm	1. No need to define number of clusters in advance. 2. Calculates a whole hierarchy of clusters. 3. Good result visualizations Joint into the methods. 4. Uses dendrogram for graphical representation	1. Inability to make corrections once the splitting/merging decision is ma 2. Lack of interpretability regarding the cluster descriptors. 3. Vagueness of termination criterion. 4. Prohibitively expensive for high dimensional and massive datasets

Figure 1: Pros and cons of clustering.

Tools and Packages

1. Tools
    - RStudio.
    - R Version 1.1.463
2. Clustering and other Packages:
    - cluster
    - factoextra (Visualisation)
    - magrittr (pipe operator in R)
    - USArrests Dataset

Dataset Description - USArrests

- This data set contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas.
- USArrests is a data frame with :
    - 50 observations(rows).
    - 4 features/variables(columns):
        * Murder : numeric, Murder arrests (per 100,000).
        * Assault: numeric, Assault arrests (per 100,000).
        * UrbanPop: numeric, Percent urban population(per 100,000).
        * Rape : numeric, Rape arrests (per 100,000).

Procedure

1. Data preparation:
    - Remove missing and junk values.
    - Scale the variables to make them equal.
2. Split the data set as (Optional):
    - Training dataset.
    - Testing dataset.
3. Implementation of clustering methodologies:
    - Computation and visualisation of k-means clustering.
    - Computation and visualisation of k-medoids clustering.
    - Computation and visualisation of PAM clustering.
    - Computation and visualisation of Hierarchial clustering.
4. Enhanced clustering visualisation using factoextra.
5. Determining optimal number of clusters.

Clustering Algorithms

- The k-Means Algorithm - Introduction:
    1. Split the input as k categories.
    2. Required: A distance measure - Eucilidean distance
    3. Processing : Clustering based on seed point computed by mean average.

- Pseudocode :

- **Initialisation:**
  - Choose a value for k.
  - Choose k random positions in the input space.
  - Assign the cluster centres μ_{j} to those positions.
- **Training**

- Repeat
  - for each datapoint x i :
    - compute the distance to each cluster centre.
    - assign the datapoint to the nearest cluster centre with distance.

$$d_{\mathrm{i}} = min_{\mathrm{j}} d(x_{\mathrm{i}}, \mu_{\mathrm{j}}). \tag{1}$$

  - for each cluster centre :
    - move the position of the centre to the mean of the points in that cluster ($\mathrm{N_j}$ is the number of points in cluster j):

$$\mu_{\mathrm{j}} = \frac{1}{N_{\mathrm{j}}} \sum_{i=1}^{N_{\mathrm{j}}} x_{\mathrm{i}} \tag{2}$$

- until the cluster centres stop moving

- **Usage**
  - for each test point :
    - Compute the distance to each cluster centre.
    - Assign the datapoint to the nearest cluster centre with distance.

$$d_{\mathrm{i}} = min_{\mathrm{j}} d(x_{\mathrm{i}}, \mu_{\mathrm{j}}). \tag{3}$$

Algorithm 1: The k-Means Algorithm

- K-medoids Algorithm - Introduction:

1. K-means is sensitive to outliers and compute large distance.
2. Change in distance measure as follows :
   - Given $x = x_1, x_2, \ldots, x_{\mathrm{N}}$
   - $median(x) \in \arg\min_{\mathrm{z}} \sum_{i=1}^{N} |x_{\mathrm{i}} - z|$
   - so could cluster by using the median to construct distance measure. More generally, we can assign cluster center to be the cluster medoid so we set $\mathrm{m_k}$:= $\mathrm{x_{i_k}}$ where

$$i_{\mathrm{k}} = \arg\min_{\mathrm{l}} \sum_{i \in C_{\mathrm{k}}} d(x_{\mathrm{l}}, x_{\mathrm{i}}). where\ l \in C_{\mathrm{k}}$$

3. Form new clusters by assigning points to the nearest cluster center.

$$C(i) = \arg\min_{\mathrm{limit}} d(x_{\mathrm{i}}, m_{\mathrm{k}}) where\ limit - 1 \leq k \leq K$$

– Pseudocode :

• **Algorithm :**

1. For a given cluster assignment C find the observation in the cluster minimizing total distance to other points in that cluster.

$$i_{\mathrm{k}} = \arg\min_{\mathrm{l}} \sum_{i \in C_{\mathrm{k}}} D(x_{\mathrm{i}}, x_{\mathrm{i}'}). \tag{4}$$

   where, $l \in C_{\mathrm{k}}, m_{\mathrm{k}} := x_{\mathrm{i_k}}$ are the current estimates of the cluster centres.

2. Given a current set of cluster centers = $\mathrm{m}_1,\ldots,\mathrm{m_K}$ minimize the total error by assigning each observation to the closest (current) cluster center:

$$C(i) = \arg\min_{\mathrm{limit}} d(x_{\mathrm{i}}, m_{\mathrm{k}}) where\ limit - 1 \leq k \leq K \tag{5}$$

3. Repeat step 1 and 2 until no change in assignment.

Algorithm 2: The k-Medoids Clustering

- Hierarcial clustering- Introduction:

1. Clusters at a given level are created by merging clusters at the next lower level.
2. Each cluster at the lowest level contains a single observation.
3. At the highest level there is just one cluster containing all the data.
4. There are two basic paradigms for constructing hierarchical clusters
   – Agglomerative : bottom-up
   – Divisive : top-down

Confusion Matrix

- A confusion matrix is a table that can be generated for a classifier on a Data Set

  True Positives(TP)- These are the cases where the predicted and actual both are yes.
  True Negatives(TN)- These are the cases where the predicted value is no and actual value is yes.
  False Positive(FP)- These are the cases where the predicted value is yes and actual value is no.
  False Negative(FN)- These are the cases where prediction is no and actual value is no.

Result

Thus the implementation of Support Vector machine is executed successfully using R program.