Cluster Analysis

Need for Clustering

• Groups available for grouping Customers

– Male and female – High Income, Medium Income, Low Income – Having average invoice amounts as < 1Lac,

between 5 Lac and 10 Lac etc. – Customers making Cheque, Cash or card payments – Any other type of grouping

• As per our convenience we categorize the entities based on some attributes

Need for Clustering

• There may be a business need of grouping the business entities on some variables.

• Such as we may group the products not just on the profit margin but also considering the variables such as quantity sold, average shelf life, maintenance cost etc.

Why is Clustering done?

• Cluster analysis is used to form groups or clusters of similar records based on several measurements made on these records.

Clustering in Marketing

• Market Segmentation: Customers are segmented based on demographic and transaction history information and a marketing strategy is tailored for each segment.

• Market structure analysis: Identifying groups of similar products according to competitive measures of similarity.

Clustering in Finance

• Creating balanced portfolios: Given data on a variety of investment opportunities (e.g., stocks), one may find clusters based on financial performance variables such as return (daily, weekly, or monthly), volatility and other characteristics, such as industry and market capitalization.

Clustering in Pure Sciences

• Biologists have made extensive use of classes and subclasses to organize species.

• In chemistry, Mendeleeyev’s periodic table of the elements.

Distance Method

• For record i we have the vector of p measurements (xi1, xi2, . . . , xip), while for record j we have the vector of measurements (xj1, xj2, . . . , xjp).

• The most popular distance measure is the Euclidean distance, dij , which between two

*d*

cases, ij =

i and j, is defined by (

*xx i*

11

− j () 2

+ xx i 22

− j (...) 2

+ + xx ip − jp ) 2

Other Distance Measures

• Numerical Data

– Correlation-based similarity – Statistical distance (also called Mahalanobis

distance) – Manhattan distance (“city block”) – Maximum coordinate distance

• Categorical Data

– Matching coefficient: (a + d)/p – Jaquard’s coefficient: d/(b+c+d)

Distance Calculation in R

• The distance can be calculated with dist( ) function in the base installation of R Syntax : dist (x , method, ...) Where

x : numeric matrix / data frame / dist object method : the distance measure to be used. Possible values are

"euclidean", "manhattan", "canberra", "binary" , "minkowski“ or

“maximum”

10

Program & Output: Distance Calculation & Scaling

11

Scaling the data

• If the variables in analysis vary in range, the variables with large values will have larger impact on the results.

• To avoid this undesirable, we should scale the data

• Variable can be scaled in the following ways:

– Subtract mean from each value(centering) and divide it by its standard

deviation(scaling) – Divide each value in the variable by maximum value of the variable – Subtract mean from each value(centering) and divide it by its mean deviation

about mean(scaling)

12

Scaling and Centering in R

• Scaling and Centering can be done by first alternative directly with function scale( ) Syntax : scale ( x , center=TRUE , scale = TRUE ) Where

x : numeric matrix center , scale : logical values

13

Steps in Clustering

1. Choosing appropriate attributes (variables) 2. Scaling and/or centering the data 3. Screening for outliers 4. Calculating Distances 5. Selecting a Cluster Algorithm 6. Obtaining one or more cluster solutions 7. Determining the number of clusters present 8. Interpreting the clusters

14

Types of Clustering Methods

• Hierarchical methods

**– Agglomerative – Divisive**

• Nonhierarchical methods

– K-Means – K-Medoids

Hierarchical Clustering

16

Distance Between Two Clusters

• Minimum distance (single linkage): the distance between the pair of records Ai and Bj that are closest.

• Maximum distance (complete linkage): the distance between the pair of records Ai and Bj that are farthest.

• Average distance (average linkage): the average distance of all possible distances between records in one cluster and records in the other cluster.

Distance Between Two Clusters

• Centroid distance: the distance between the two cluster centroids. A cluster centroid is the vector of measurement averages across all the records in that cluster.

• Ward: ANOVA sum of squares between the two clusters added up over the variables

Agglomerative Hierarchical Method

• Agglomerative method begins with n (No. of observations) clusters and sequentially merge similar clusters until a single cluster is left.

Algorithm

1. Start with n clusters (each observation =

cluster). 2. The two closest observations are merged into

one cluster. 3. At every step, the two clusters with the

smallest distance are merged. This means that either single observations are added to existing clusters or two existing clusters are combined.

Hierarchical Clustering in R

• Hierarchical Clustering in R can be implemented using function hclust( ) Syntax : hclust ( d , method , ...) Where

d : dissimilarity structure produced by dist () distance function method : can values as “ward.D”(default), “single”, “complete”,

“average”, etc.

21

Example: Milk (dataset in package flexclust)

• The data set contains the ingredients of mammal's milk of 25 animals.

• A data frame with 25 observations on the following 5 variables (all in percent)

– water – protein – fat – lactose – ash

• Here, we are interested in grouping the 25 mammals based on the above given 5 nutrient measures

22

Program & Output: Dendogram

23

Dendrogram

• Dendrogram displays how items are combined into clusters and is read from the bottom up.

• Each observation starts as its own cluster.

• Two the observations that are closest are combined

• Then the clustered observations are combined with the closest further

• This continues until all observations are combined into a single cluster

24

Dendrogram

Height indicates the criterion value at which clusters are joined. For average linking, this criterion is the average distance between each point in one cluster and each point in another cluster

25

NbClust Package

• NbClust package offers many indices for finding the best number of clusters. Function NbClust ( ) is provided.

• The indices might not be consistent with each other, but the results can be a guide for selecting the optimal value of number of clusters Syntax : NbClust(data, distance, min.nc = 2, max.nc = 15, method, ...) Where

data : data frame / matrix distance : Distance method

min.nc : minimum number of clusters max.nc : maximum number of clusters to be tried method : method of linkage of clusters

26

Program & Output : NBClust( )

27

Output of Graphs

28

Program & Output: Best No. of Clusters

• Best.nc object gives the Best number of clusters proposed by each index and the corresponding index value.

29

Program & Output: Best No. of Clusters

• Hence we extract the clustering indices along with best number of clusters proposed by each and plot it.

30

We see here that the maximum number of votes are for cluster number 3. Hence we recommend the number of clusters to be 3 for further analysis

Obtaining Clusters

• The cluster IDs for observations can be obtained with the function cutree() Syntax : cutree ( tree , k ) Where

tree : object produced by hclust() k : Cluster number used to cut the tree

31

Program & Output : Cluster IDs

32

Program & Output

33

Limitations of Hierarchical Clustering

• Hierarchical clustering requires the computation and storage of an n × n distance matrix. For very large datasets, this can be expensive and slow.

• The hierarchical algorithm makes only one pass through the data. This means that records that are allocated incorrectly early in the process cannot be reallocated subsequently.

Limitations of Hierarchical Clustering

• With respect to the choice of distance between clusters, single and complete linkage are robust to changes in the distance metric (e.g., Euclidean, statistical distance) as long as the relative ordering is kept.

• Hierarchical clustering is sensitive to outliers.

K-Means Clustering

36

K-Means

• Forming good clusters is to pre-specify a desired number of clusters, k, and to assign each case to one of k clusters so as to minimize a measure of dispersion (variation) within the clusters.

• The method divides the sample into a predetermined number k of non-overlapping clusters so that clusters are as homogeneous as possible with respect to the measurements used.

Algorithm 1. Start with k initial clusters (user chooses k).

The starting points are chosen by software 2. At every step, each record is reassigned to the

cluster with the “closest” centroid. 3. Re-compute the centroids of clusters that lost

or gained a record, and repeat step 2. 4. Stop when moving any more records

between clusters increases cluster dispersion.

K-means in R

• K-means can be implemented in R by function kmeans( ) in stat package. Syntax : kmeans ( x , centers , ...) Where

x : numeric matrix of data, or an object that can be coerced

to such a matrix centers : number of clusters

39

Program & Output

40

Within Sum of Squares

• Within Sum of Squares indicate the variation

• We can create a function to extract and plot it.

• Alternatively we can also use NbClust() function

• Lower the WSS better is the cluster variation.

• WSS can be extracted from the component tot.withinSS of kmeans object.

41

Program & Output

42

Advantages of K-Means Clutering

• K-means clustering can handle larger datasets than hierarchical cluster approach.

• Observations are not permanently committed to any cluster but, they are changed at every iteration.

43

Limitations

• All the variables have to be continuous / numeric

• Clusters are severely affected by outliers

• Clusters are sensitive to initialization

• Clusters obtained are of differing densities

44