

CS 422

DATA MINING

**HOMEWORK ASSIGNMENT 3**  
PRACTICUM QUESTIONS 1 AND 2

RONIT RUDRA  
A20379221  
rrudra@hawk.iit.edu

Course Instructor  
Prof. Jawahar PANCHAL

## 1 Question 1:

- 1.1 Load the auto-mpg sample dataset into the Orange application - ensure that origin is set as a target attribute type, as it will be used as a class label. Perform a Hierarchical Clustering using Linkage set to Average , after calculating Distances, with Pruning set to a Max Depth of 5. Also, set Selection to Top N with a value of 3. This will result in a shallow tree of depth 5, and a nal cut resulting in 3 clusters. Examine the resulting clusters (C1,C2,C3) via Distributions analysis - is there a clear relationship between the cluster assignment and class label (1,2,3)? What are the probabilities calculated for each value of origin for each cluster? Does changing the Max Depth affect the results in any way?

### 1.2 Answer:

- Loading the dataset and changing *origin* to a target variable.

Figure 1: Load Data File

File

☒ File:

☐ URL:

Info

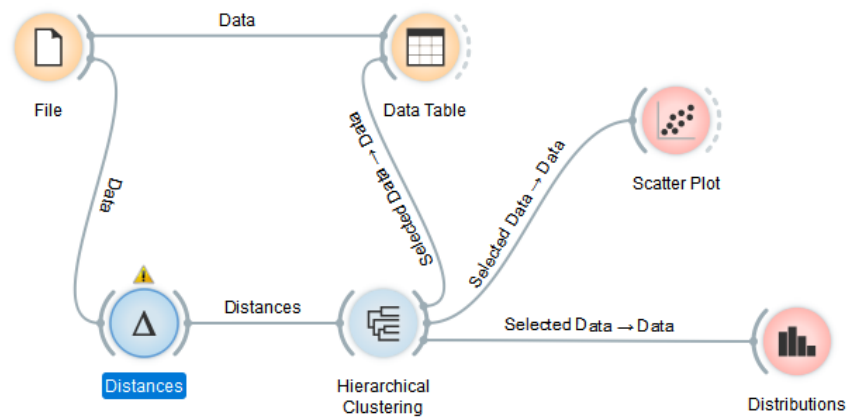
398 instance(s), 8 feature(s), 0 meta attribute(s)  
Regression; numerical class.

Columns (Double click to edit)

1	cylinders	<span>D</span> nominal	feature	3, 4, 5, 6, 8
2	displacement	<span>C</span> numeric	feature	
3	horsepower	<span>C</span> numeric	feature	
4	weight	<span>C</span> numeric	feature	
5	acceleration	<span>C</span> numeric	feature	
6	model_year	<span>D</span> nominal	feature	70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82
7	origin	<span>D</span> nominal	target	1, 2, 3
8	car_name	<span>D</span> nominal	feature	amc ambassador brougham, amc ambassador dpl, amc ambassador sst, amc concord, amc ...
9	mpg	<span>C</span> numeric	feature	

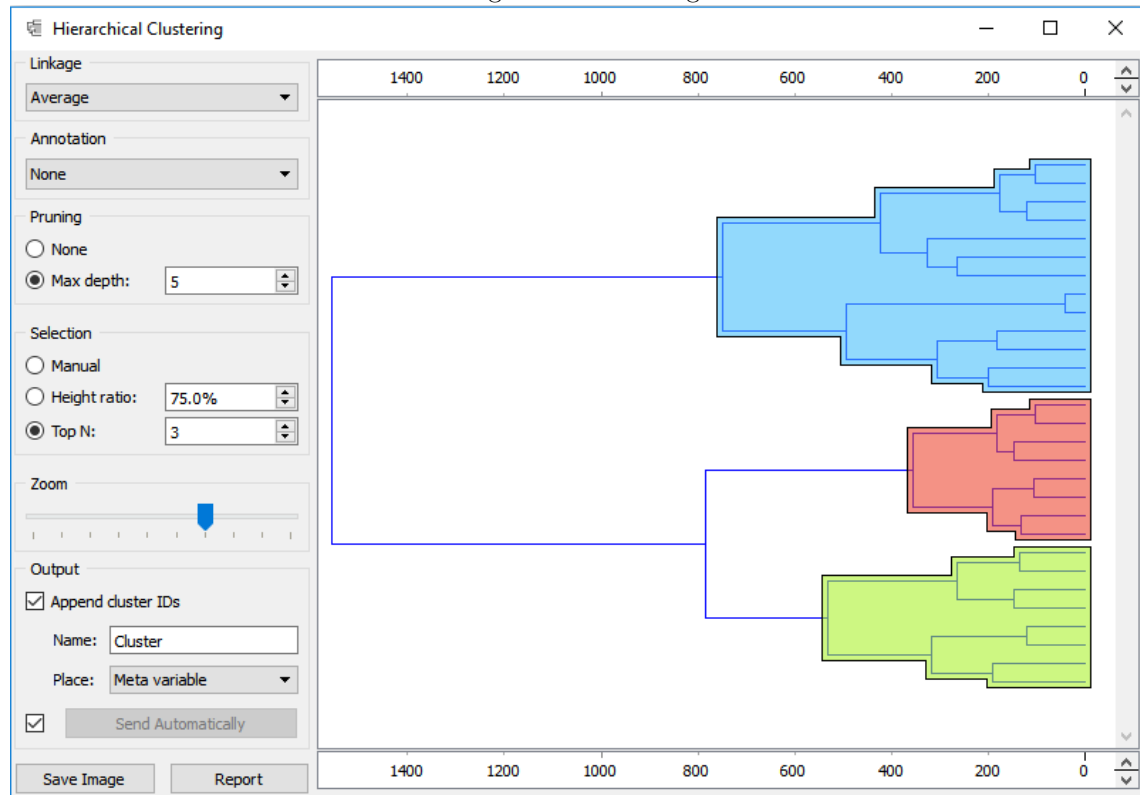
- Creating the appropriate workflow. The euclidean distance metric is calculated and then passed on to the Heirarchial Clustering widget.

Figure 2: Workflow



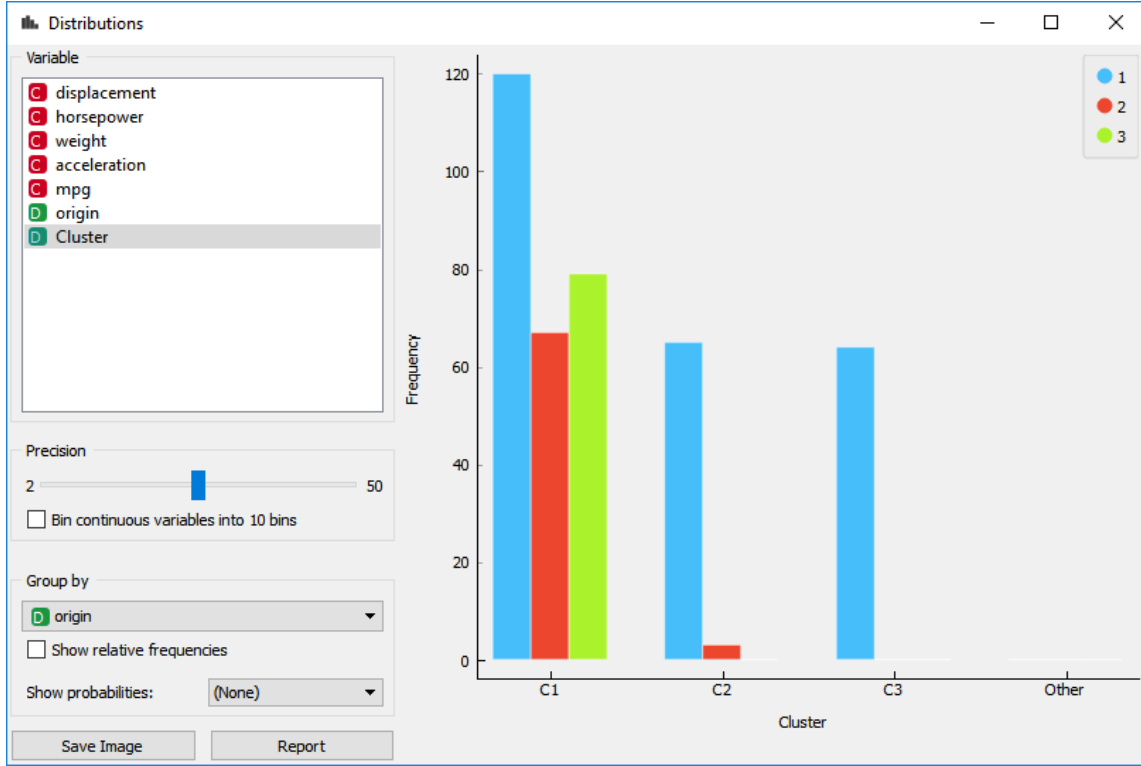
- Setting up parameters of Hierarchical Clustering.

Figure 3: Clustering



- Distribution Analysis.

Figure 4: Distribution Plots of Formed Clusters.



- Including the Probabilities tab we get the probability of each class belonging to a cluster.

Table 1: Class Probabilities

Cluster	Origin 1	Origin 2	Origin 3
1	0.451	0.252	0.297
2	0.956	0.044	0
3	1.0	0.0	0.0

### 1.3 Observations:

On observation of Figure 4. it is seen that class 2 and 3 are clustered together while class 1 is spread out among all the three clusters. There does not seem to be a definitive relation among the three classes except the fact that class 2 and 3 seem similar. The probability distribution is shown in Figure 5. Changing the *max depth* does not result in a change, only if it is equal to 1 does the number of clusters reduce to 2 which is pretty obvious.

## 2 Question 2:

2.1 Load the breast-cancer-wisconsin-cont dataset into the Orange application, and run a k-means analysis with the number of clusters Optimized From values for k from 2 to 5. Use Silhouette scoring - what is the score for each value of k? For the best score, what are the coordinates of the centroids? What are the distances between the centroids for the best score?

### 2.2 Answer:

- Loading the dataset.

Figure 5: Load Data File

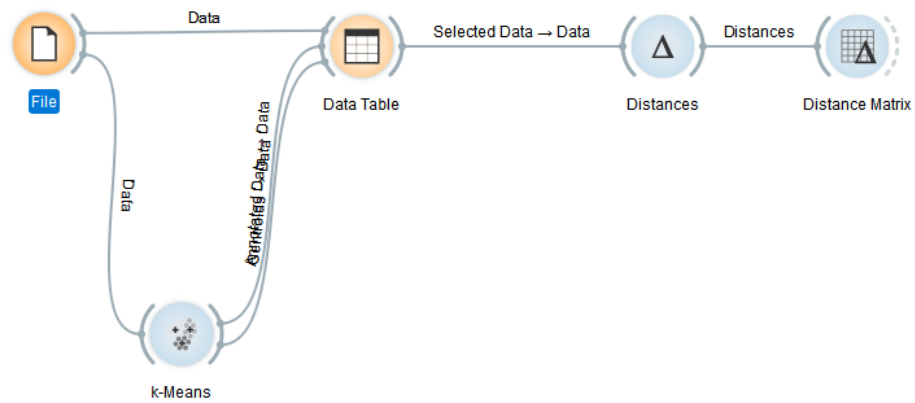
The screenshot shows the 'File' widget in the Orange data mining software. The 'File' radio button is selected, and the file 'breast-cancer-wisconsin-cont.tab' is loaded. The 'Info' section indicates 683 instances, 9 features, and 0 meta-attributes, with a classification target having 2 discrete values. The 'Columns' section lists 10 columns, with the last column, 'type', highlighted as the target variable.

Column Index	Column Name	Icon	Data Type	Role	Values
1	Clump thickness	C	numeric	feature	
2	Unif_Cell_Size	C	numeric	feature	
3	Unif_Cell_Shape	C	numeric	feature	
4	Marginal_Adhe...	C	numeric	feature	
5	Single_Cell_Size	C	numeric	feature	
6	Bare_Nuclei	C	numeric	feature	
7	Bland_Chromat...	C	numeric	feature	
8	Normal_Nucleoli	C	numeric	feature	
9	Mitoses	C	numeric	feature	
10	type	D	nominal	target	benign, malign

Buttons at the bottom: 'Browse documentation data sets' and 'Report'.

- Creating the appropriate workflow.

Figure 6: Workflow



- Performing K-Means Clustering with the given parameters.



Figure 7: K-Means Clustering

**Number of Clusters**

☐ Fixed: 8

☒ Optimized from 2 to 5

Scoring: Silhouette

**Initialization**

Initialize with KMeans++

Re-runs: 10

Maximal iterations: 300

**Output**

Append cluster ID as: Class

Name: Cluster

Report ☒ Apply Automatically

**Scoring (bigger is better)**

k	Score
2	0.59
3	0.52
4	0.48
5	0.23

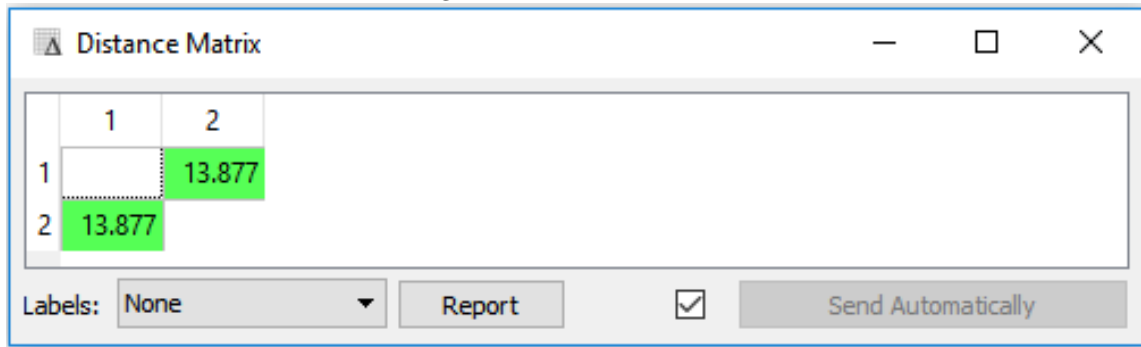
- Passing Centroid Coordinates of model with highest silhouette score to display as table.

Figure 8: Centroid Coordinates

	Clump_Thickness	Unit_Cell_Size	Unit_Cell_Shape	Marginal_Adhesion	Single_Cell_Size	Bare_Nuclei	Band_Chromatin	Normal_Nucleoli	Mitoses
1	6.786	6.360	6.283	5.236	4.988	7.509	5.634	5.541	2.158
2	2.597	0.805	0.946	0.844	1.619	0.849	1.696	0.781	0.620

- Calculating euclidean distance between centroids using distance widget and visualized by distance matrix widget.

Figure 9: Centroid Distance



### 2.3 Observations:

For models with cluster numbers ranging from 2 to 5, the silhouette scores are shown in Figure 8. with model with  $k=2$  wins out with a score of 0.59 . For this model, the centroid coordinates are shown in Figure 9. For the same model, the distance between the centroids is shown in figure 10. and is calculated to be 13.87 .