Data warehouse and Data Mining (TDT 4300)

Assignment 3

Submitted by

Name : Md Anwarul Hasan
Student ID : 583233

Task 1 k-Means Clustering

Answer:

I have tried this one, but could not finish it. My program is giving me error. I have attached my code also with the assignment.

Task 2 Hierarchical Agglomerative Clustering (HAC)

(a) Explain the Hierarchical Agglomerative Clustering (HAC) and the difference between MINlink and MAX-link.

Answer:

First, HAC considers every point in the data as a different cluster. Then, the two most close cluster were merged. When the merger is done then there is one less cluster exist in the system. For example, if there are n points then there are n clusters and after the 1^{st} cluster formation between two points (clusters), based on their distance property, there are (n-1) cluster left in the data. So on and so forth this process continues until all the points (clusters) come into one single cluster. Other than the initial data points that forms initial cluster with most closest data points, if there is other points then they may have different “proximmities” or distances between them. By such a way the MAX-link and MIN-link are present in HAC.

In MIN-link two clusters are merges if it is found that, they are the closest datapoints / have lowest distance between them. Cluster consists of multiple points need to checked for all the points to confirm the minimum distance property.

In MAX-link two clusters are merged if it is found that, between them the distance is maximum. Cluster consists of multiple points need to checked for all the points to confirm the maximum distance property.

(b) You are given a two-dimensional dataset shown in Table 1. Perform HAC (for both MINlink and MAX-link) and present the results in the form of dendrogram. Use the Euclidean distance. Describe thoroughly the process and the outcome of each step.

ID	*x*	*y*
A	5	4
B	4	7
C	6	8
D	8	2
E	12	7
F	11	6

Answer:

The process is given below. The distance between points are calculated by Euclidian distance calculation method.

Distance A-B = Sqrt $((5-4)^2 + (4-7)^2)$ = sqrt (1 + 9) = 3.16

Likewise,

Distance A-C = Sqrt $((5-6)^2 + (4-8)^2)$ = sqrt (1 + 16) = 4.12

Distance A-D = Sqrt $((5-8)^2 + (4-2)^2)$ = sqrt (9 + 4) = 3.60

Distance A-E = Sqrt $((5-12)^2 + (4-7)^2)$ = sqrt (49 + 9) = 7.62

Distance A-F = Sqrt $((5-11)^2 + (4-6)^2)$ = sqrt (36 + 4) = 6.32

Distance B-C = Sqrt $((4-6)^2 + (7-8)^2)$ = 2.24

Distance B-D = Sqrt $((4-8)^2 + (7-2)^2)$ = 6.40

Distance B-E = Sqrt $((4-12)^2 + (7-7)^2)$ = 8.0

Distance B-F = Sqrt $((4-11)^2 + (7-6)^2)$ = 7.07

Distance C-D = Sqrt $((6-8)^2 + (8-2)^2)$ = 6.32

Distance C-E = Sqrt $((6-12)^2 + (8-7)^2)$ = 6.08

Distance C-F = Sqrt $((6-11)^2 + (8-6)^2)$ = 5.39

Distance D-E = Sqrt $((8-12)^2 + (2-7)^2)$ = 6.40

Distance D-F = Sqrt $((8-11)^2 + (2-6)^2)$ = 5.0

Distance E-F = Sqrt $((12-11)^2 + (7-6)^2) = 1.41$

So the Distance matrix is,

	A	B	C	D	E	F
A	0	3.16	4.12	3.60	7.62	6.32
B		0	2.24	6.40	8.0	7.07
C			0	6.32	6.08	5.39
D				0	6.40	5.0
E					0	1.41
F						0

Min-Link:

i) lowest distance is E-F, looks at all distances from the other data points to nodes E and F, and compares in pairs which of the two choices will give Min-Link, i.e. shortest distance between data points in the clusters.

Min {(A,E), (A,F)} = (A,F)

Min {(B,E),(B,F)} = (B,F)

Min {(C,E), (C,F)} = (C,F)

Min {(D,E), (D,F)} = (D,F)

ii) Now we update the table with the updated local-best cluster distances from all other nodes to our new cluster {E, F} cluster. Now see that the lowest distance is B-C. do the same as for the previous point for the {E, F} cluster

	A	B	C	D	{E, F}
A	0	3.16	4.12	3.60	6.32
B		0	2.24	6.40	7.07
C			0	6.32	5.39
D				0	5.0
{E, F}					0

Min {(A,B), (A,C)} = (A,B)

Min {(D,B), (D,C)} = (D,C)

Min {({E, F},B), ({E, F},C)} = ({E, F},C)

iii) From now onwards we continue doing this until every point or cluster falls or comes under single cluster

	A	{B, C}	D	{E, F}
A	0	3.16	3.60	6.32
{B, C}		0	6.32	5.39
D			0	5.0
{E, F}				0

iv) Now, {A, (B, C)} is the shortest and updating the table we get,

	{A, {B, C}}	D	{E, F}
{A, {B, C}}	0	3.60	5.39
D		0	5.0
{E, F}			0

v) Now, {{A, (B, C)}, D} is we have left {{A, (B, C)}, D} and {E, F}

vi) Now, we have every cluster under one cluster.

Max-Link:

i) After clustering the most closest points/ clusters, i.e E and F, now we find the most distant points from all other points to the points E and F to form our new distance matrix.

Max {(A,E), (A,F)} = (A,F)
Max {(B,E),(B,F)} = (B,F)
Max {(C,E), (C,F)} = (C,F)
Max {(D,E), (D,F)} = (D,F)

	A	B	C	D	{E, F}
A	0	3.16	4.12	3.60	7.62
B		0	2.24	6.40	8.0
C			0	6.32	6.08
D				0	6.40
{E, F}					0

And we continue this procedure until all the points comes under single cluster.

ii) Now, distance B, C is lowest. Merging them and calculating the next distance matrix, we get,

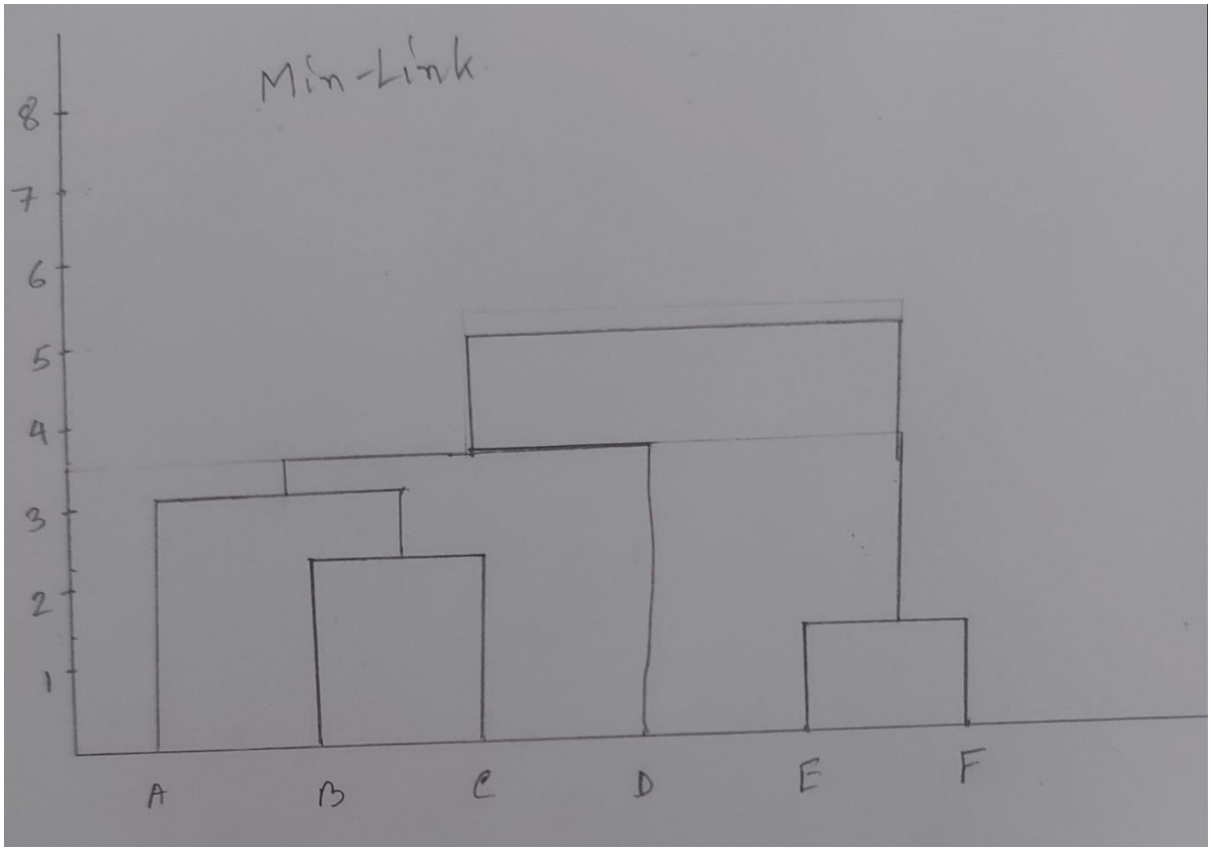
	A	{B, C}	D	{E, F}
A	0	4.12	3.60	7.62
{B, C}		0	6.40	8.0
D			0	6.40
{E, F}				0

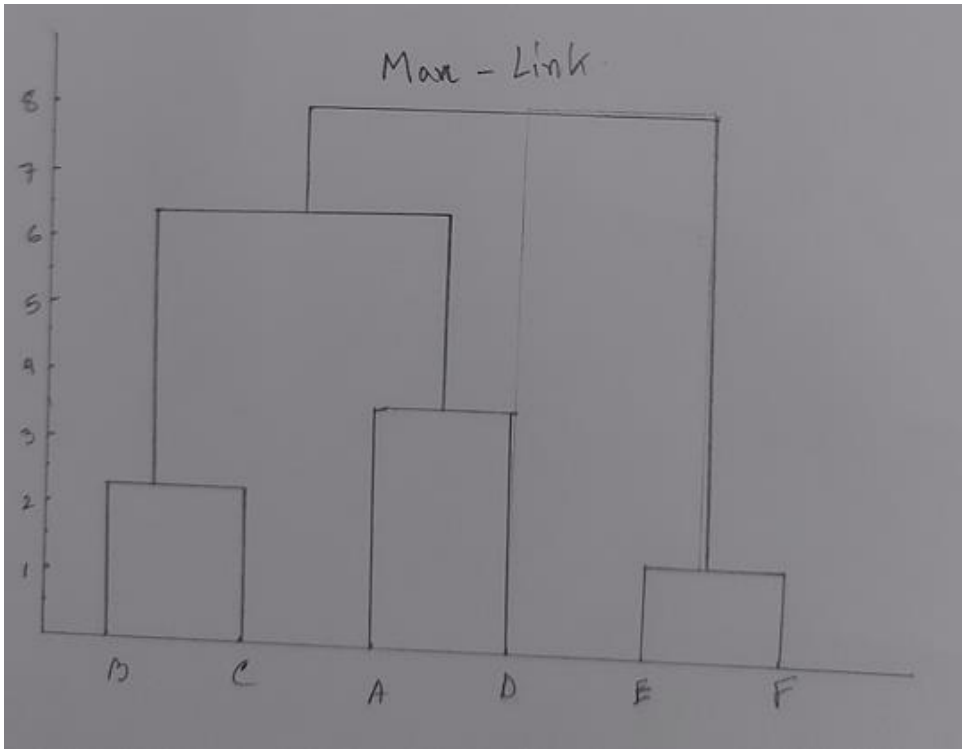
iii) Now, distance A, D is lowest. Merging them and calculating the next distance matrix, we get,

	{A, D}	{B, C}	{E, F}
{A, D}	0	6.40	7.62
{B, C}		0	8.0
{E, F}			0

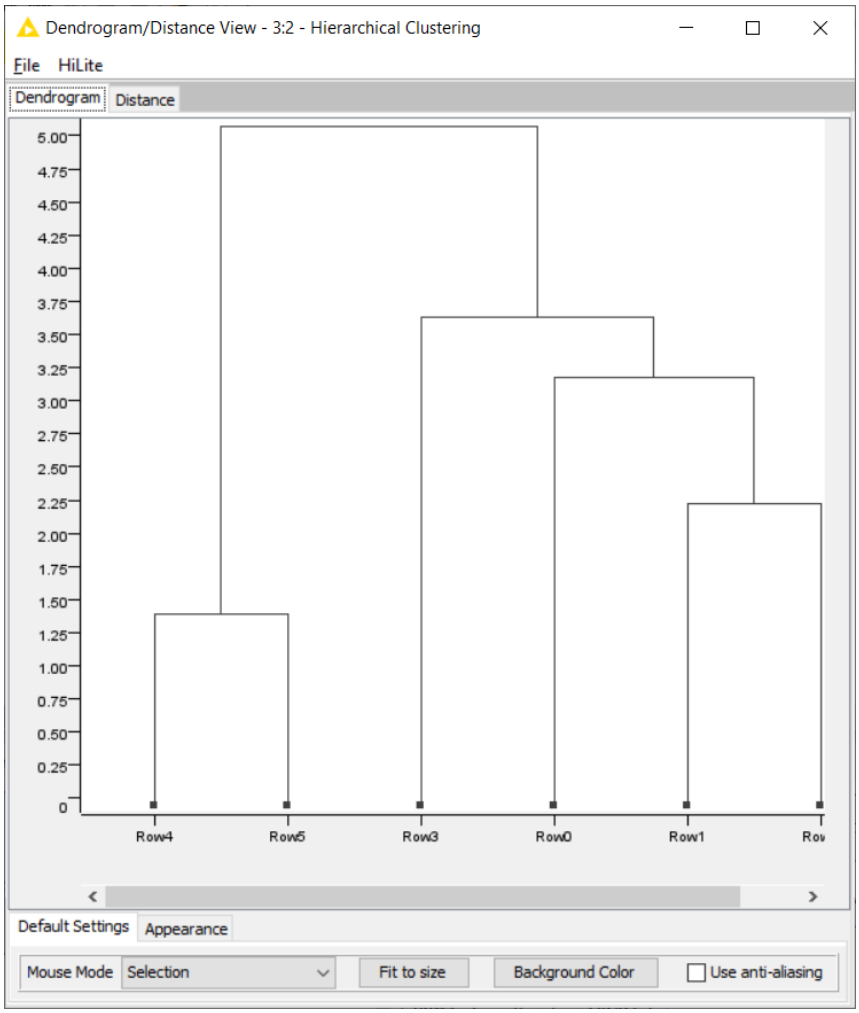
iv) Now, distance { A, D}, {B, C} is lowest. And so we have now {{ A, D}, {B, C} } and {E, F}.

v) At this stage every point is under a single cluster.

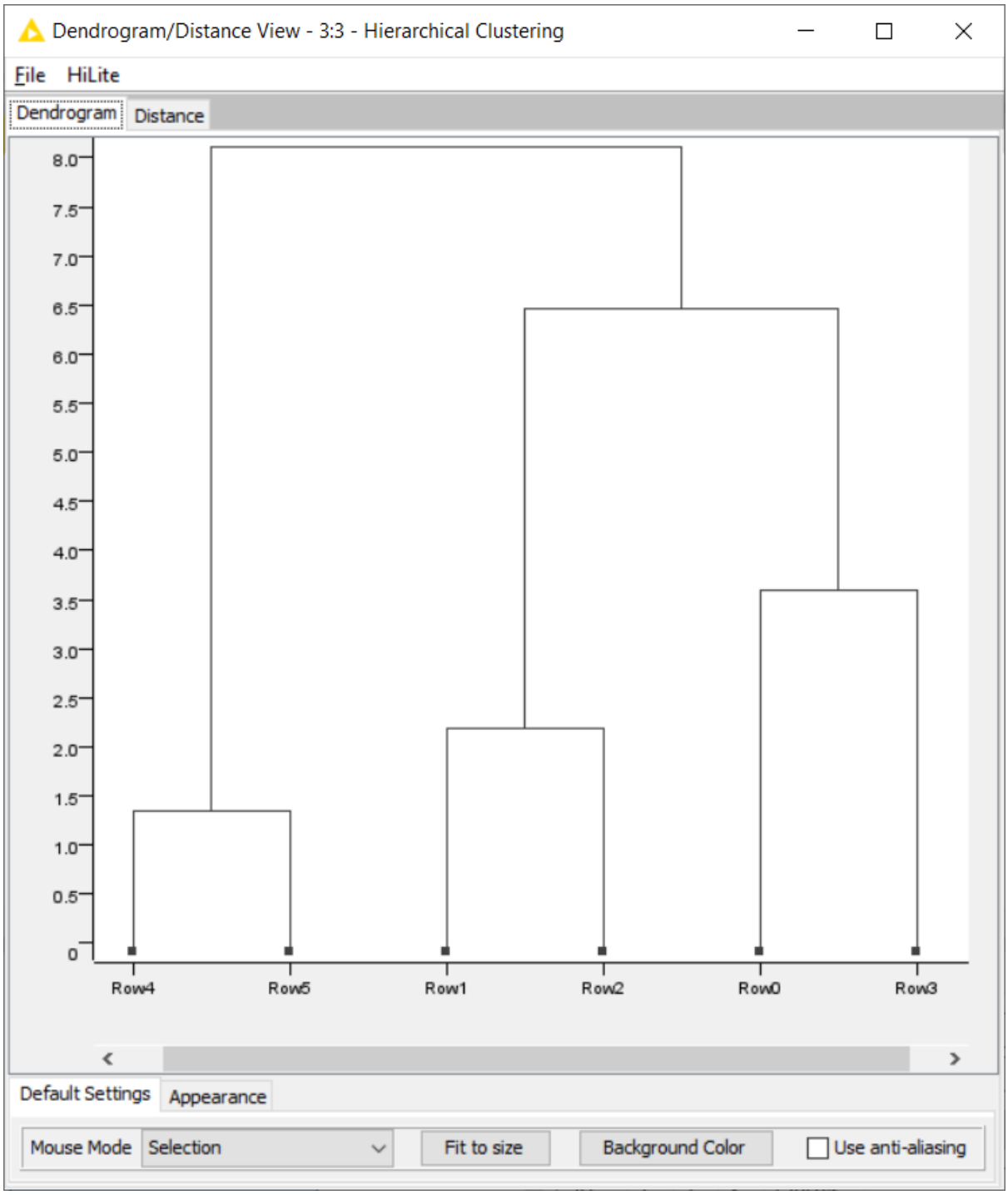



Max - Link
8
7
6
5
4
3
2
1
B
C
A
D
E
F

Min-link:


Dendrogram/Distance View - 3:2 - Hierarchical Clustering
File HiLite
Dendrogram Distance
5.00
4.75
4.50
4.25
4.00
3.75
3.50
3.25
3.00
2.75
2.50
2.25
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0
Row4
Row5
Row3
Row0
Row1
Default Settings Appearance
Mouse Mode Selection
Fit to size
Background Color
Use anti-aliasing

Max-Link:


Dendrogram/Distance View - 3:3 - Hierarchical Clustering
File
HiLite
Dendrogram
Distance
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0
Row4
Row5
Row1
Row2
Row0
Row3
Default Settings
Appearance
Mouse Mode
Selection
Fit to size
Background Color
Use anti-aliasing

Task 2 DBSCAN Clustering

You are given following points: P1 = (2;2), P2 = (13;7), P3 = (7;13), P4 = (2;1), P5 = (6;12), P6 = (12;5), P7 = (3;11), P8 = (13;9), P9 = (1;2), P10 = (9;2), P11 = (4;8), P12 = (9;11), P13 = (15;6), P14 = (3;5), P15 = (7;5), P16 = (3;2), P17 = (12;6), P18 = (12;12).

(a) Your task is to perform DBSCAN clustering given the parameters Eps = 3 (Euclidean metric) and MinPts = 2 (including the analyzed point). Identify core, border and noise points. Identify clusters. **Describe thoroughly the process and the outcome of each step.**

Answer:

Provided points are,

Points	X	Y
P1	2	2
P2	13	7
P3	7	13
P4	2	1
P5	6	12
P6	12	5
P7	3	11
P8	13	9
P9	1	2
P10	9	2
P11	4	8
P12	9	11
P13	15	6
P14	3	5
P15	7	5
P16	3	2
P17	12	6
P18	12	12

Also provided that, Eps = 3 and MinPts = 2

So, to calculate core, border and noise points we need to first create the distance matrix between points with the help of Euclidian distance,

For example, distance between points P1 and P2 will be calculated like below,

P1 – P2 distancce = $\text{sqrt}((2-13)^2 + (2-7)^2) = \text{sqrt}(121 + 25) = 12.08$

I created the distance matrix in MS Excel and presenting that below,

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18
P1	-																	
P2	12.08	-																
P3	12.08	8.49	-															
P4	1.00	12.53	13.00	-														
P5	10.77	8.60	1.41	11.70	-													
P6	10.44	2.24	9.43	10.77	9.22	-												
P7	9.06	10.77	4.47	10.05	3.16	10.82	-											
P8	13.04	2.00	7.21	13.60	7.62	4.12	10.20	-										
P9	1.00	13.00	12.53	1.41	11.18	11.40	9.22	13.89	-									
P10	7.00	6.40	11.18	7.07	10.44	4.24	10.82	8.06	8.00	-								
P11	6.32	9.06	5.83	7.28	4.47	8.54	3.16	9.06	6.71	7.81	-							
P12	11.40	5.66	2.83	12.21	3.16	6.71	6.00	4.47	12.04	9.00	5.83	-						
P13	13.60	2.24	10.63	13.93	10.82	3.16	13.00	3.61	14.56	7.21	11.18	7.81	-					
P14	3.16	10.20	8.94	4.12	7.62	9.00	6.00	10.77	3.61	6.71	3.16	8.49	12.04	-				
P15	5.83	6.32	8.00	6.40	7.07	5.00	7.21	7.21	6.71	3.61	4.24	6.32	8.06	4.00	-			
P16	1.00	11.18	11.70	1.41	10.44	9.49	9.00	12.21	2.00	6.00	6.08	10.82	12.65	3.00	5.00	-		
P17	10.77	1.41	8.60	11.18	8.49	1.00	10.30	3.16	11.70	5.00	8.25	5.83	3.00	9.06	5.10	9.85	-	
P18	14.14	5.10	5.10	14.87	6.00	7.00	9.06	3.16	14.87	10.44	8.94	3.16	6.71	11.40	8.60	13.45	6.00	-

So points with the neighbours with the boundary of radius Eps = 3 are,

Points	**Neighbours**
P1	P4, P9, P16
P2	P6, P8, P13, P17
P3	P5, P12
P4	P1, P9, P16
P5	P3,
P6	P2, P17
P7	-
P8	P2
P9	P1, P4, P16
P10	-
P11	-
P12	P3
P13	P2, P17
P14	P16
P15	-
P16	P1, P4, P9, P14
P17	P2, P6, P13
P18	-

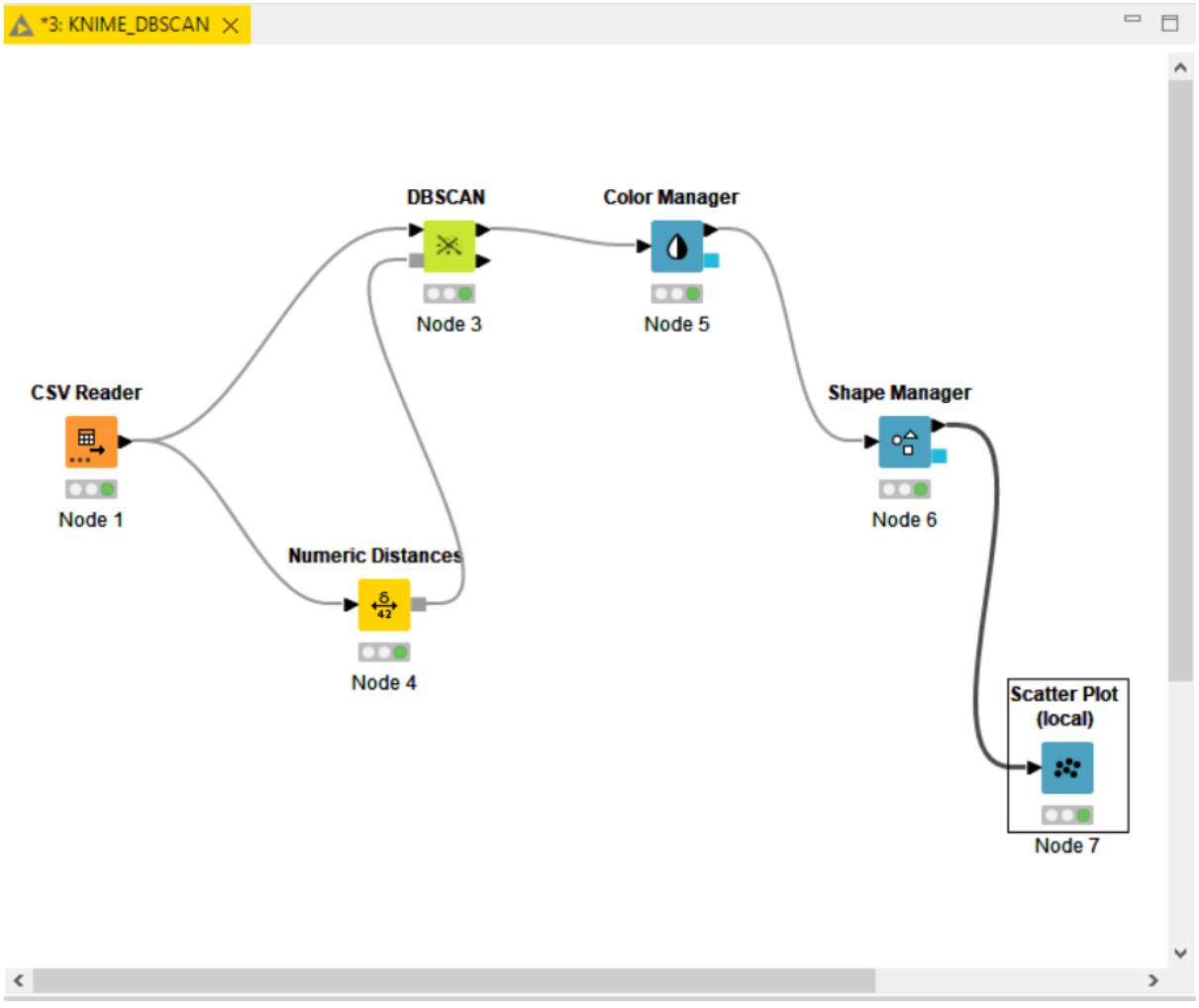
As the MinPts is 2, then core points are: {P1, P2, P3, P4, P5, P6, P8, P9, P12, P13, P14, P16, P17}

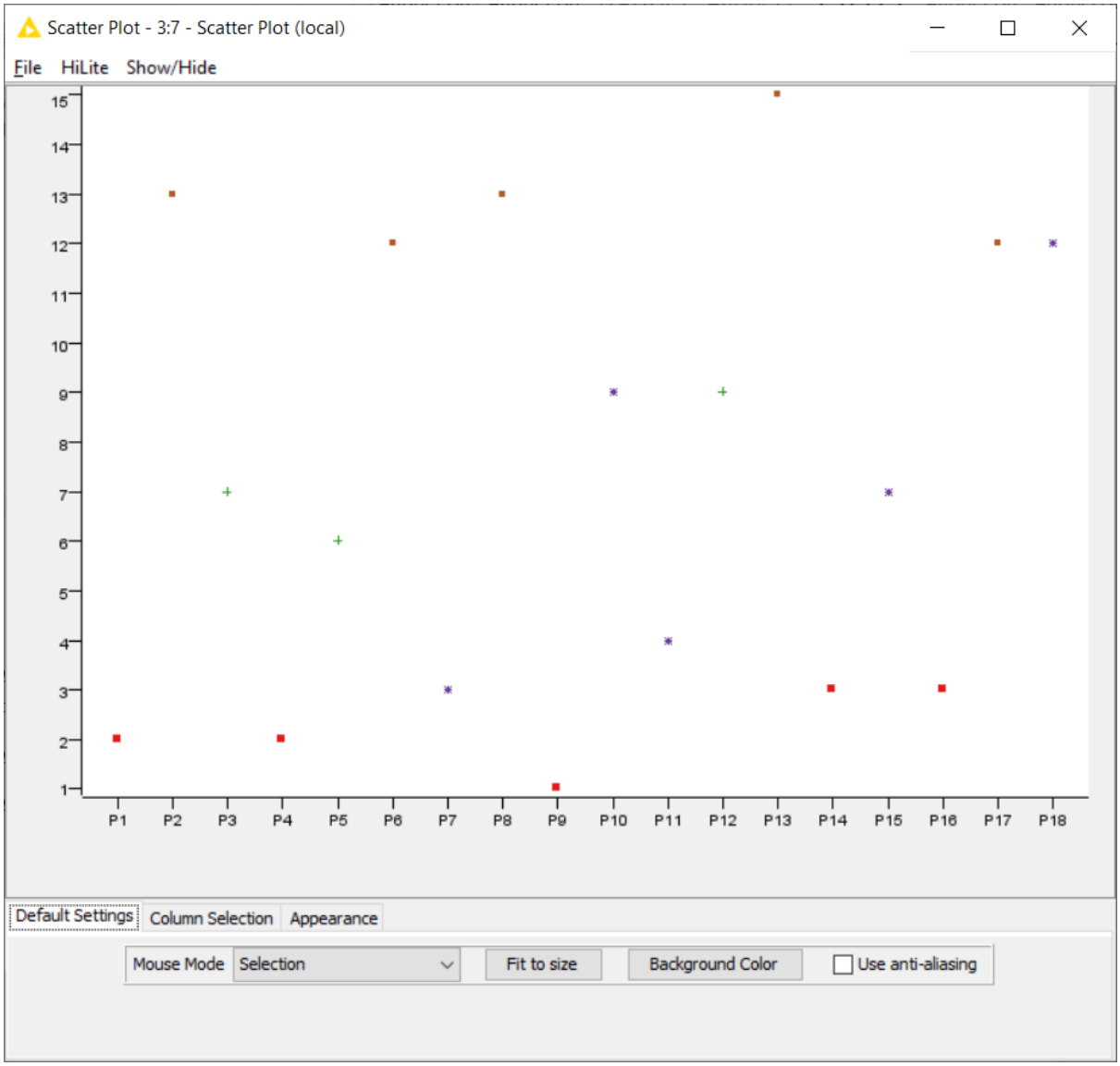
And the border points are: { Ø}

And noise points are: {P7, P10, P11, P15, P18}

(b) Verify your results using the KNIME data analytics platform. We provide you the file ata_dbscan.csv containing the very same data. **Present a picture of your workflow and the scatter plot with marked clusters and outliers.**

Answer:




Scatter Plot - 3:7 - Scatter Plot (local)
File
HiLite
Show/Hide
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
P1
P2
P3
P4
P5
P6
P7
P8
P9
P10
P11
P12
P13
P14
P15
P16
P17
P18
Default Settings
Column Selection
Appearance
Mouse Mode
Selection
Fit to size
Background Color
Use anti-aliasing