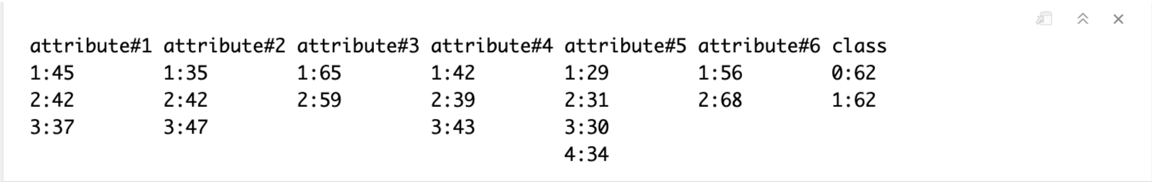
Advanced Data Mining Assignment 3: Association Analysis 2

Nazli Bilgic (nazbi056) & Simge Cinar (simci637)

March 5^{th} 2024

1 Introduction

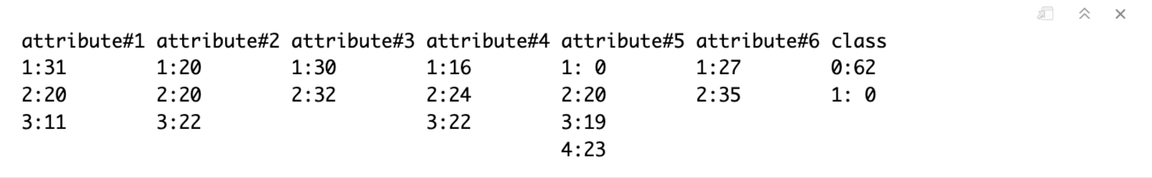
The monk1 dataset has 6 categorical attributes with binary class labels. The dataset has 124 instances. The summary table can be seen below.



```
attribute#1 attribute#2 attribute#3 attribute#4 attribute#5 attribute#6 class
1:45        1:35        1:65        1:42        1:29        1:56        0:62
2:42        2:42        2:59        2:39        2:31        2:68        1:62
3:37        3:47                    3:43        3:30
                                                4:34
```

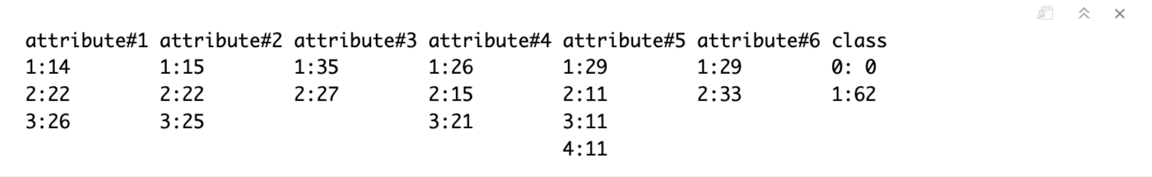
Figure 1: summary table for the dataset

The summary table for class 1 and class 0 is as follows:



```
attribute#1 attribute#2 attribute#3 attribute#4 attribute#5 attribute#6 class
1:31        1:20        1:30        1:16        1: 0        1:27        0:62
2:20        2:20        2:32        2:24        2:20        2:35        1: 0
3:11        3:22                    3:22        3:19
                                                4:23
```

Figure 2: summary table for class 0

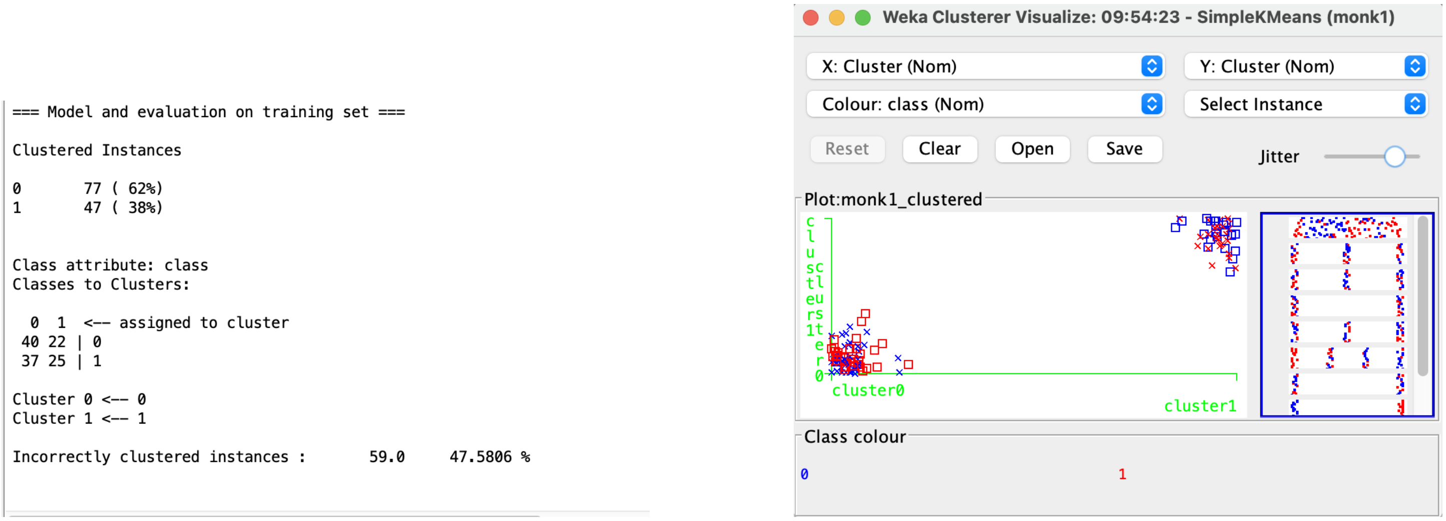


```
attribute#1 attribute#2 attribute#3 attribute#4 attribute#5 attribute#6 class
1:14        1:15        1:35        1:26        1:29        1:29        0: 0
2:22        2:22        2:27        2:15        2:11        2:33        1:62
3:26        3:25                    3:21        3:11
                                                4:11
```

Figure 3: summary table for class 1

2 Clustering

First, k-means clustering is applied with where k = 2 (seed = 10). Incorrectly clustered instances is 59 (47.5807%). The algorithm performed poorly hence accuracy is low (Figure 4).



```
=== Model and evaluation on training set ===

Clustered Instances

0       77 ( 62%)
1       47 ( 38%)


Class attribute: class
Classes to Clusters:

  0  1  <-- assigned to cluster
 40 22 | 0
 37 25 | 1

Cluster 0 <-- 0
Cluster 1 <-- 1

Incorrectly clustered instances :       59.0     47.5806 %
```

Weka Clusterer Visualize: 09:54:23 - SimpleKMeans (monk1)

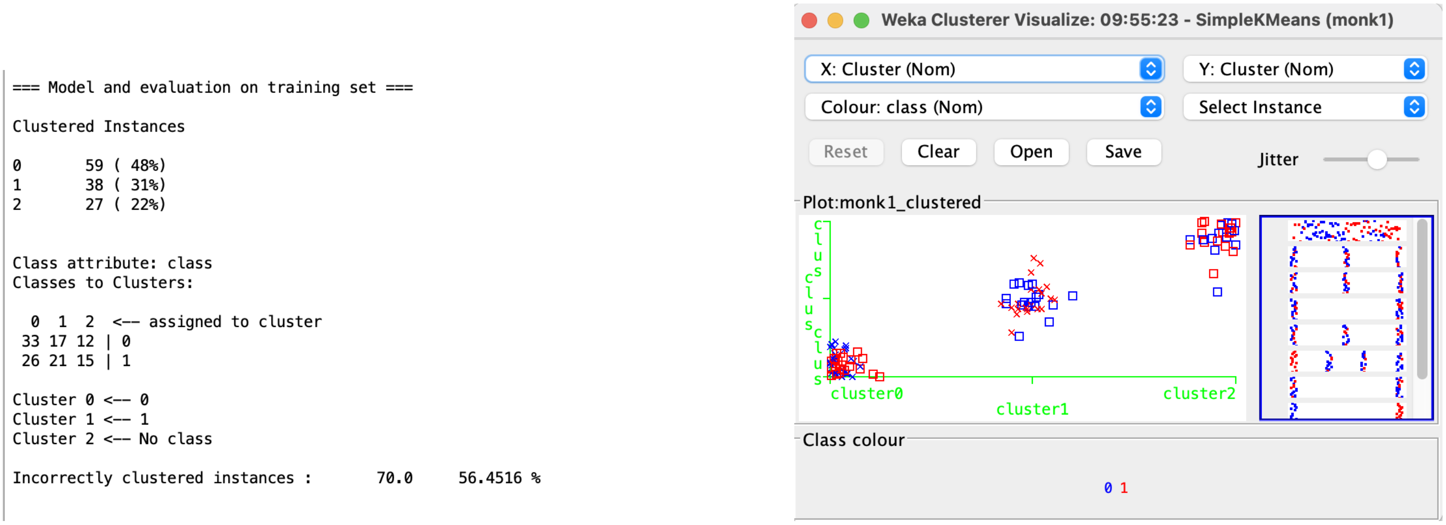
X: Cluster (Nom) Y: Cluster (Nom)

Colour: class (Nom)

(a) Output (b) Visualisation of the clusters vs class

Figure 4: k-means with 2 cluster output

k-means clustering is applied with where k = 3 (seed = 10). Incorrectly clustered instances is 70 (56.4516%), accuracy is low (Figure 5).



```
=== Model and evaluation on training set ===

Clustered Instances

0       59 ( 48%)
1       38 ( 31%)
2       27 ( 22%)


Class attribute: class
Classes to Clusters:

  0  1  2  <-- assigned to cluster
 33 17 12 | 0
 26 21 15 | 1

Cluster 0 <-- 0
Cluster 1 <-- 1
Cluster 2 <-- No class

Incorrectly clustered instances :       70.0     56.4516 %
```

Weka Clusterer Visualize: 09:55:23 - SimpleKMeans (monk1)

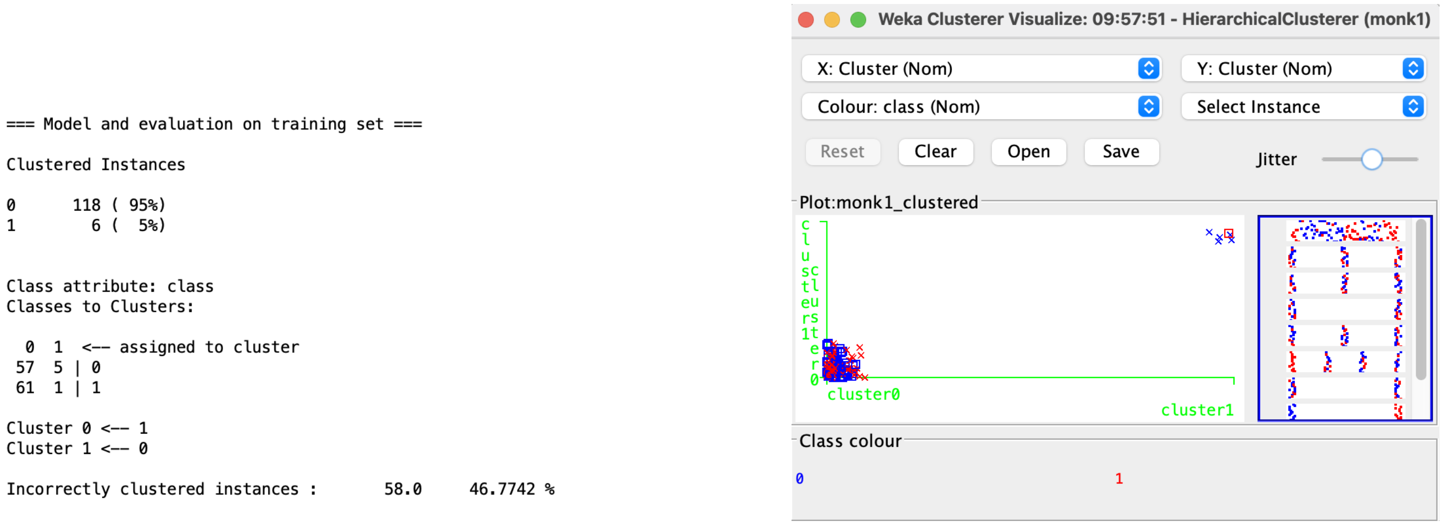
X: Cluster (Nom) Y: Cluster (Nom)

Colour: class (Nom)

(a) Output (b) Visualisation of the clusters vs class

Figure 5: k-means with 3 cluster output

Then HierarchicalClusterer is applied is with 2 clusters where 'linkType = COMPLETE'. Most of the instances assigned to cluster 0. Incorrectly clustered instances is 58 (46.7742%) and accuracy is low.



```
=== Model and evaluation on training set ===

Clustered Instances

0      118 ( 95%)
1        6 (  5%)


Class attribute: class
Classes to Clusters:

  0  1  <-- assigned to cluster
 57  5 | 0
 61  1 | 1

Cluster 0 <-- 1
Cluster 1 <-- 0

Incorrectly clustered instances :       58.0     46.7742 %
```

Weka Clusterer Visualize: 09:57:51 - HierarchicalClusterer (monk1)

X: Cluster (Nom)

Y: Cluster (Nom)

Colour: class (Nom)

Select Instance

Plot:monk1_clustered

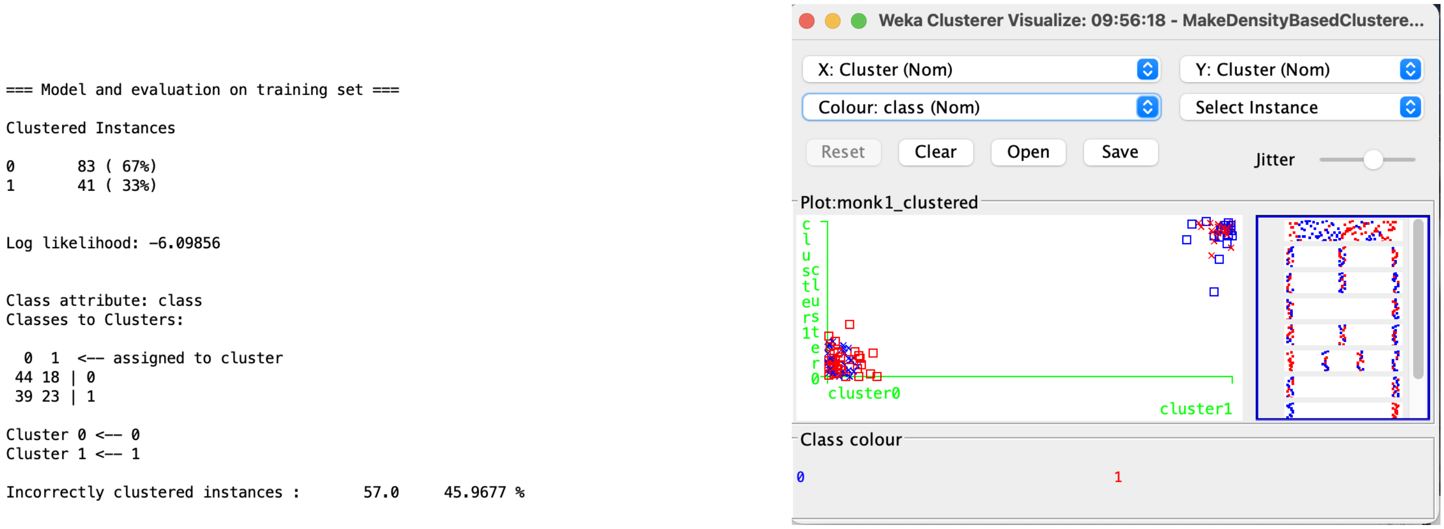
Class colour

(a) Output

(b) Visualisation of the clusters vs class

Figure 6: Hierarchical clustering with 2 clusters

Finally, MakeDensityBasedClusterer is applied with 2 clusters and 0.01 minimum standard deviation. Incorrectly clustered instances is 57 (45.9677%) and accuracy is low again.



```
=== Model and evaluation on training set ===

Clustered Instances

0       83 ( 67%)
1       41 ( 33%)


Log likelihood: -6.09856


Class attribute: class
Classes to Clusters:

  0  1  <-- assigned to cluster
 44 18 | 0
 39 23 | 1

Cluster 0 <-- 0
Cluster 1 <-- 1

Incorrectly clustered instances :       57.0     45.9677 %
```

Weka Clusterer Visualize: 09:56:18 - MakeDensityBasedClustere...

X: Cluster (Nom)

Y: Cluster (Nom)

Colour: class (Nom)

Select Instance

Plot:monk1_clustered

Class colour

(a) Output

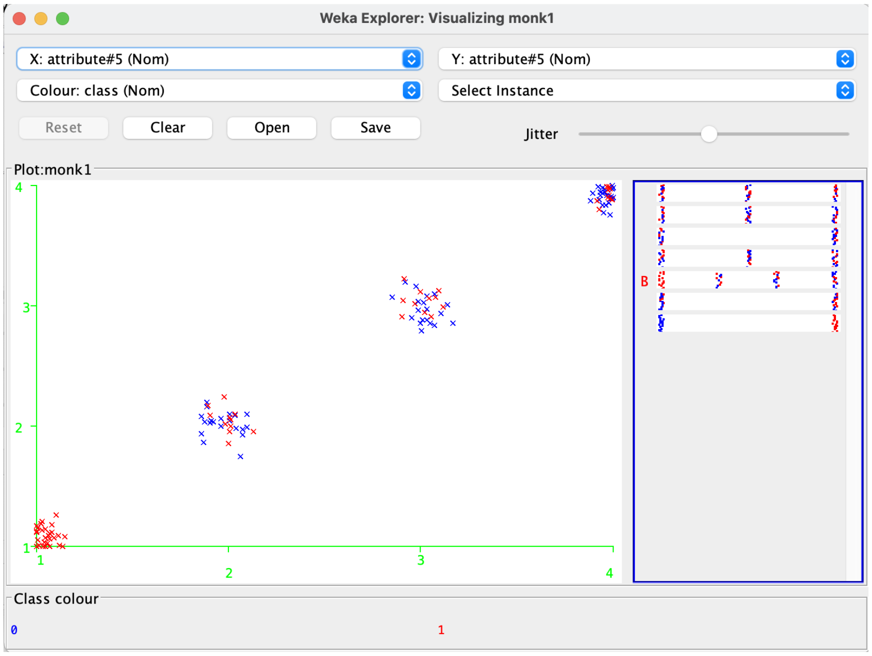
(b) Visualisation of the clusters vs class

Figure 7: MakeDensityBasedClusterer with 2 clusters

All of the attributes are categorical and this might be the reason for low accuracy in the clustering algorithms. Clustering algorithms typically rely on distance measures or similarity metrics, which are not as straightforward to define for categorical attributes hence it caused low performance.

3 Association Analysis

The scatterplots for some attributes can be seen below. First of all, when attribute#5 = 1, the class is 1 . Secondly, when attribute#1 and attribute#2 have the same category, class is 1. These statements can be observed from Figure 8 and Figure 9 respectively.



Weka Explorer: Visualizing monk1

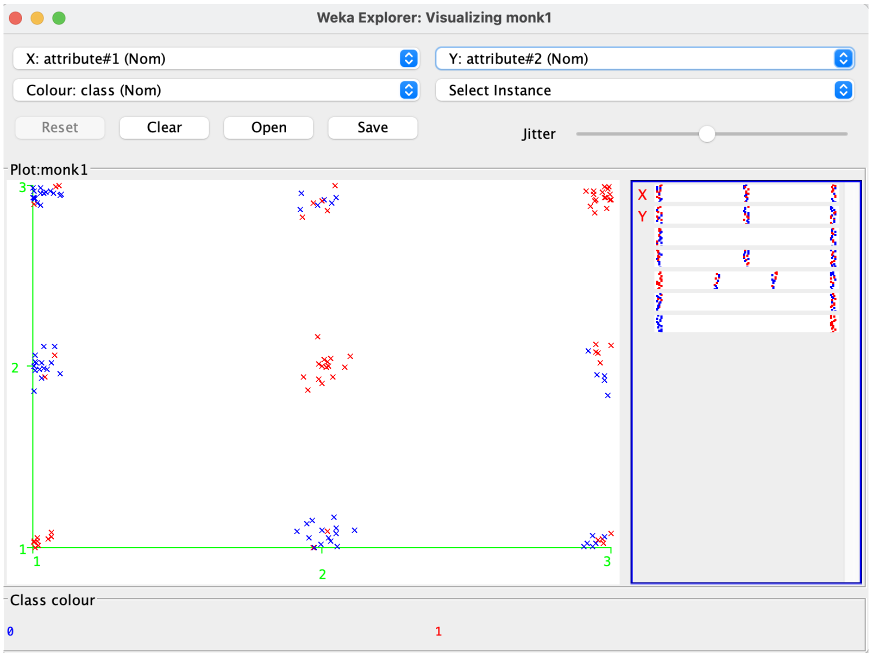
X: attribute#5 (Nom)

Y: attribute#5 (Nom)

Colour: class (Nom)

Select Instance

Figure 8: attribute#5 scatterplot



Weka Explorer: Visualizing monk1

X: attribute#1 (Nom)

Y: attribute#2 (Nom)

Colour: class (Nom)

Select Instance

Figure 9: attribute#1 vs attribute#2 scatterplot

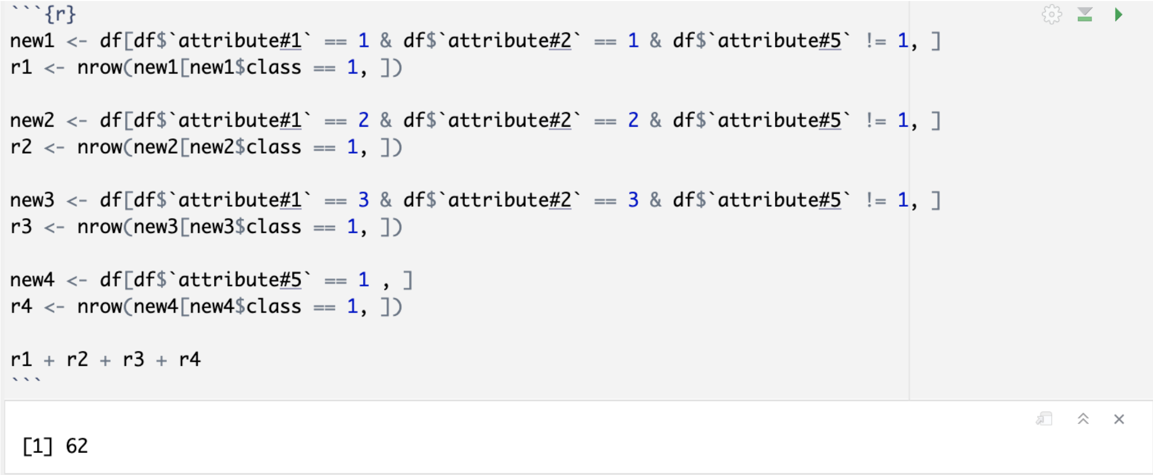
The four rules below determine the class 1. Since class is binary, it is enough to find rules for only one class. The calculations using R can be seen in the Figure 6 below. There are 62 instances in class 1 and when the dataset is filtered according to each rule it sums up 62.

$$\text{Rule 1: } \{\text{attribute\#5} = 1\}\ (29) \rightarrow \{\text{class} = 1\}\ (29)\ \text{conf:}(1) \tag{1}$$

$$\text{Rule 2: } \{\text{attribute\#1} = 3, \text{attribute\#2} = 3\}\ (17) \rightarrow \{\text{class} = 1\}\ (17)\ \text{conf:}(1) \tag{2}$$

$$\text{Rule 5: } \{\text{attribute\#1} = 2, \text{attribute\#2} = 2\}\ (15) \rightarrow \{\text{class} = 1\}\ (15)\ \text{conf:}(1) \tag{3}$$

$$\text{Rule 14: } \{\text{attribute\#1} = 1, \text{attribute\#2} = 1\}\ (9) \rightarrow \{\text{class} = 1\}\ (9)\ \text{conf:}(1) \tag{4}$$



~~~
```{r}
new1 <- df[df$`attribute#1` == 1 & df$`attribute#2` == 1 & df$`attribute#5` != 1, ]
r1 <- nrow(new1[new1$class == 1, ])

new2 <- df[df$`attribute#1` == 2 & df$`attribute#2` == 2 & df$`attribute#5` != 1, ]
r2 <- nrow(new2[new2$class == 1, ])

new3 <- df[df$`attribute#1` == 3 & df$`attribute#2` == 3 & df$`attribute#5` != 1, ]
r3 <- nrow(new3[new3$class == 1, ])

new4 <- df[df$`attribute#5` == 1 , ]
r4 <- nrow(new4[new4$class == 1, ])

r1 + r2 + r3 + r4
```
~~~

```
[1] 62
```

Figure 10: Calculations in R

The mixed instances within clusters, as observed in the outputs in Section 2, suggest that the class labels are influenced by complex attribute interactions and cannot captured by simple distance or density measures. Clustering, as an unsupervised approach, does not leverage the pre-defined labels that are integral to the MONK1 dataset.