lab3 data mining

Tore Andersson , Zahra Jalilpour

March 2021

Data Description:
At first we analysis the data set. Monk1 data set consists of 124 instances and following six nominal attributes and one class which is binary target attribute as seen in table 1: In figure 1 shows a sample of how the data is coded in the monk1 dataset.

X-i	name	Characteristic
x1	Head-Shape	round, square,octagon
x2	Body-Shape	round, square, octagon
x3	Is-smiling	yes, no
x4	Holding	sword, balloon, flag
x5	Jacket-color	red,yellow, green, blue
x6	Has-tie	yes, no

Table 1: Attribute code, Attribute name, Attribute space

Relation: monk1

No.	attribute#1 Nominal	attribute#2 Nominal	attribute#3 Nominal	attribute#4 Nominal	attribute#5 Nominal	attribute#6 Nominal	class Nominal
1	1	1	1	1	3	1	1
2	1	1	1	1	3	2	1
3	1	1	1	3	2	1	1
4	1	1	1	3	3	2	1
5	1	1	2	1	2	1	1
6	1	1	2	1	2	2	1
7	1	1	2	2	3	1	1
8	1	1	2	2	4	1	1
9	1	1	2	3	1	2	1
10	1	2	1	1	1	2	1

Figure 1: Sample of instances and attributes based on class

From Figure 2 we can see one of the reasons for why the data could be hard to cluster. When all the variables are nominal and there is plenty of overlapping between the two classes for same values for most attributes paired against each other and it is clear that attribute 1 and attribute 2 are dependent to each other.

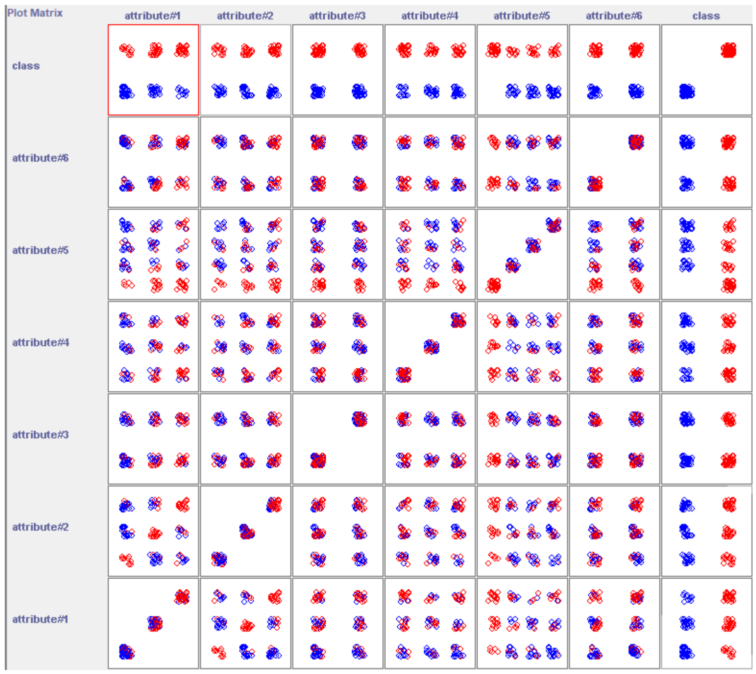

Figure 2: Visualization of the data set

Different clustering methods

All clustering were created with seed = 10. We want to analysis clustering in this data set. We have applied different clustering algorithms. Simple Kmeans clustering, EM clustering, Density based clustering and Hierachical clustering by different number of cluster. As we see in table 2, the number of cluster does not have any effect in the output. In all different algorithms,number of Incorrectly cluster instances are high.

By analysing this data set through different algorithms, we found that Hierachical clustering performed better than other kinds of clustering to specify the classes of each instances. We have tried this algorithm by three different values of k, but in k=3 and k=5 the number of incorrectly clustered instances are lower.

In k=5, we can see the number of instances in three clusters are one.So it could not be a good clustering. In k=3, one cluster has one instance and other instances are distributed equally in two clusters. So we consider this clustering which can be seen in figure 3 for associated analysis. Also the accuracy is low, but it is better than all other algorithms that we tried earlier. For rest of the clustering outputs see Appendix 1.

Clustering model	K	Incorrectly clustered: num	:%
Simple K means	2	59	47.5
	3	70	56.5
	5	79	63.7
Hiearchical clustering	2	61	49.1
	3	48	38.7
	5	48	38.7
Density based	2	57	45.9
	3	66	53.2
	4	78	62.9
EM	x	53	42.7

Table 2: Table of tested clustering models

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

Clustered Instances

0       67 ( 54%)
1       56 ( 45%)
2        1 (  1%)


Class attribute: class
Classes to Clusters:

  0  1  2  <-- assigned to cluster
 41 21  0 | 0
 26 35  1 | 1

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

Incorrectly clustered instances :       48.0     38.7097 %
```

Figure 3: Hierachical clustering K = 3

Association Analysis

For associated analysis, we use Hierachical clustering by k=3. We add the cluster attribute to the dataset. By considering number of rules=19 and minimum support=0.05, we see the best rules as below.

Now we should do some filtering to consider the rules that not containing antecedent. and as want to predict the cluster, we should consider the rules which have cluster as output in figure 4.

Before analysing association analysis, we can see, as the instances are categorical variables, clustering algorithms that uses Euclidean and Manhatan distances will not be a good clustering for our data set, because these kind of clustering is used

```
Minimum support: 0.05 (6 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 19

Generated sets of large itemsets:

Size of set of large itemsets L(1): 19

Size of set of large itemsets L(2): 151

Size of set of large itemsets L(3): 378

Size of set of large itemsets L(4): 125

Size of set of large itemsets L(5): 6

Best rules found:

 1. attribute#5=1 29 ==> class=1 29    conf:(1)
 2. attribute#1=3 attribute#2=3 17 ==> class=1 17    conf:(1)
 3. attribute#3=1 attribute#5=1 17 ==> class=1 17    conf:(1)
 4. attribute#5=1 attribute#6=1 16 ==> class=1 16    conf:(1)
 5. attribute#1=2 attribute#2=2 15 ==> class=1 15    conf:(1)
 6. attribute#1=3 attribute#5=1 13 ==> class=1 13    conf:(1)
 7. attribute#5=1 attribute#6=2 13 ==> class=1 13    conf:(1)
 8. attribute#2=3 attribute#5=1 12 ==> class=1 12    conf:(1)
 9. attribute#3=2 attribute#5=1 12 ==> class=1 12    conf:(1)
10. attribute#1=3 attribute#2=3 attribute#6=2 12 ==> class=1 12    conf:(1)
11. attribute#4=1 attribute#5=1 11 ==> class=1 11    conf:(1)
12. attribute#1=2 attribute#5=1 10 ==> class=1 10    conf:(1)
13. attribute#2=2 attribute#5=1 10 ==> class=1 10    conf:(1)
14. attribute#1=1 attribute#2=1 9 ==> class=1 9    conf:(1)
15. attribute#4=2 attribute#5=1 9 ==> class=1 9    conf:(1)
16. attribute#4=3 attribute#5=1 9 ==> class=1 9    conf:(1)
17. attribute#1=2 attribute#2=2 attribute#3=1 9 ==> class=1 9    conf:(1)
18. attribute#1=3 attribute#2=3 attribute#3=1 9 ==> class=1 9    conf:(1)
19. attribute#3=1 attribute#5=1 attribute#6=1 9 ==> class=1 9    conf:(1)
```

Figure 4: Associated analysis by Hierachical clustering with k = 3

for numerical features. It is the main reason that accuracy in confusion matrix is low.
Now we should find as few rules predicting class 1 as possible. Here we should delete redundant rules, by viewing description of each attributes, it is clear that attribute 1 and 2 are related to each other. Here we select the rules with max confidence and logically related to each other.

- attribute5 =1 29 class=1 29 conf:(1) (rule1 Jacket color=red)
- attribute1 =3 attribute2 =3 17 conf:(1) (rule2 head and body shape is octagon)
- attribute1 =2 attribute2 =2 15 conf:(1) (rule5 head and body shape is

square)

- attribute1 =1 attribute2 =1 9 conf:(1) (rule14 head and body shape is round)

As we can see, the four selected rules have max confidence which each instances in each class are logically related to each other.Therefore we can conclude a-priori algorithm can classified this dataset by considering the structure of instances.

Appendix: 1

```
=== 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 %
```

(a) K = 2

```
=== 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 %
```

(b) K =3

Figure 5: Simple Kmeans clustering with K= 2,3

From 5

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

Clustered Instances

0       39 ( 31%)
1       34 ( 27%)
2       22 ( 18%)
3       12 ( 10%)
4       17 ( 14%)


Class attribute: class
Classes to Clusters:

  0  1  2  3  4  <-- assigned to cluster
 26 15  9  3  9 | 0
 13 19 13  9  8 | 1

Cluster 0 <-- 0
Cluster 1 <-- 1
Cluster 2 <-- No class
Cluster 3 <-- No class
Cluster 4 <-- No class

Incorrectly clustered instances :       79.0     63.7097 %
```

(a) K = 5

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

Clustered Instances

0       59 ( 48%)
1       65 ( 52%)


Log likelihood: -6.00606


Class attribute: class
Classes to Clusters:

  0  1  <-- assigned to cluster
 34 28 | 0
 25 37 | 1

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

Incorrectly clustered instances :       53.0     42.7419 %
```

(b) EM clustering

Figure 6: Simple Kmeans clustering with K=5 and EM clustering

```
=== 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 %
```

(a) K = 2

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

Clustered Instances

0       60 ( 48%)
1       39 ( 31%)
2       25 ( 20%)


Log likelihood: -6.09108


Class attribute: class
Classes to Clusters:

  0  1  2  <-- assigned to cluster
 35 16 11 | 0
 25 23 14 | 1

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

Incorrectly clustered instances :       66.0     53.2258 %
```

(b) K = 3 clustering

Figure 7: Density based clustering with K = 2,3

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

Clustered Instances

0       51 ( 41%)
1       35 ( 28%)
2       23 ( 19%)
3       15 ( 12%)


Log likelihood: -6.06035


Class attribute: class
Classes to Clusters:

  0  1  2  3  <-- assigned to cluster
 28 17 12  5 | 0
 23 18 11 10 | 1

Cluster 0 <-- 0
Cluster 1 <-- 1
Cluster 2 <-- No class
Cluster 3 <-- No class

Incorrectly clustered instances :       78.0     62.9032 %
```

(a) Density based, K = 4

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

Clustered Instances

0      123 ( 99%)
1        1 (  1%)


Class attribute: class
Classes to Clusters:

  0  1  <-- assigned to cluster
 62  0 | 0
 61  1 | 1

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

Incorrectly clustered instances :       61.0     49.1935 %
```

(b) Hierarchical, K = 2

Figure 8: Density based clustering with K = 4 and Hierachical clustering K = 2

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

Clustered Instances

0       67 ( 54%)
1       56 ( 45%)
2        1 (  1%)


Class attribute: class
Classes to Clusters:

  0  1  2  <-- assigned to cluster
 41 21  0 | 0
 26 35  1 | 1

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

Incorrectly clustered instances :       48.0     38.7097 %
```

(a) K = 3

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

Clustered Instances

0       65 ( 52%)
1       56 ( 45%)
2        1 (  1%)
3        1 (  1%)
4        1 (  1%)


Class attribute: class
Classes to Clusters:

  0  1  2  3  4  <-- assigned to cluster
 41 21  0  0  0 | 0
 24 35  1  1  1 | 1

Cluster 0 <-- 0
Cluster 1 <-- 1
Cluster 2 <-- No class
Cluster 3 <-- No class
Cluster 4 <-- No class

Incorrectly clustered instances :       48.0     38.7097 %
```

(b) K = 4

Figure 9: Hierarchical clustering K = 3,4