732A75 Data Mining Lab-2

Anubhav Dikshit(anudi287) and Nahid Farazmand (nahfa911)

21 Feburary 2019

Dataset

In this exercise we want to cluster a given dataset and use association analysis to describe the clusters obtained. for this purpose, we will work with one of the most well-known datasets in the data mining literature, namely the Iris dataset. The dataset consists of 50 samples from each of three species of Iris flowers (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from each sample, they are the length and the width of sepal and petal.

Clustering

Here we know there are 3 types of Iris flowers; so K is 3 and we have chosen the seed = 10 (from the previous exercise we know that changing the rate of seed will change the initial starting points of the K-mean process).

The following parameters were chosen:

- Minimum support = 0.1
- Metric type: confidence
- Minimum confidence = 0.9

K-Means

The results with k-means are quite good with 3 bins and 3 clusters. From the confusion matrix we can see that all the iris setosa items are classified correctly, while versicolor and virginica sometimes get swapped.

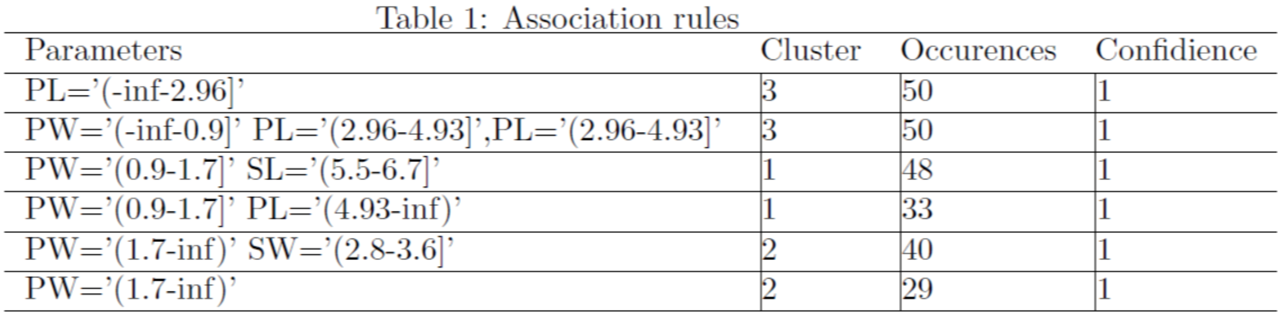


Table 1: Association rules

Parameters	Cluster	Occurences	Confidience
PL='(-inf-2.96]'	3	50	1
PW='(-inf-0.9]' PL='(2.96-4.93]',PL='(2.96-4.93]'	3	50	1
PW='(0.9-1.7]' SL='(5.5-6.7]'	1	48	1
PW='(0.9-1.7]' PL='(4.93-inf)'	1	33	1
PW='(1.7-inf)' SW='(2.8-3.6]'	2	40	1
PW='(1.7-inf)'	2	29	1

```
Class attribute: class
Classes to Clusters:

  0  1  2  <— assigned to cluster
  0  0 50 | Iris-setosa
 48  2  0 | Iris-versicolor
  7 43  0 | Iris-virginica

Cluster 0 <— Iris-versicolor
Cluster 1 <— Iris-virginica
Cluster 2 <— Iris-setosa

Incorrectly clustered instances :      9.0      6      %
```

Association analysis and Visualization

By using the default settings for the Apriori algorithm, we do not get enough rules to cover all clusters. Instead we used numRules = 1000, metricType = confidence and minMetric = 0.9. In this way we obtain many rules and we decided to keep rules which have more than 90% confidence. The following are the interesting rules that we got.

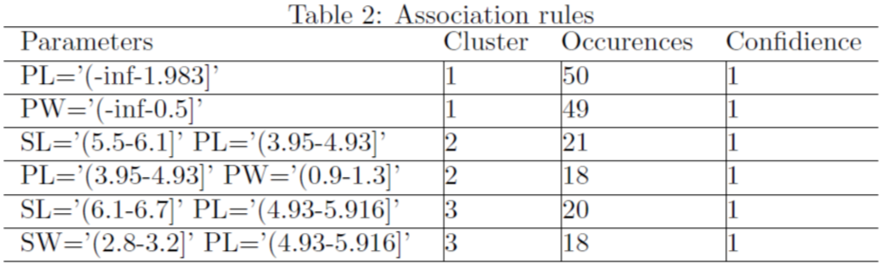


Table 2: Association rules

Parameters	Cluster	Occurences	Confidience
PL='(-inf-1.983]'	1	50	1
PW='(-inf-0.5]'	1	49	1
SL='(5.5-6.1]' PL='(3.95-4.93]'	2	21	1
PL='(3.95-4.93]' PW='(0.9-1.3]'	2	18	1
SL='(6.1-6.7]' PL='(4.93-5.916]'	3	20	1
SW='(2.8-3.2]' PL='(4.93-5.916]'	3	18	1

The reason that we do not want to have the class attribute in the antecedent is that it is what we want to predict to begin with, so when we have a new flower to consider we do not have its class at our disposal. Moreover we only want the cluster attribute in the consequent, because the other attributes we want to use to predict.

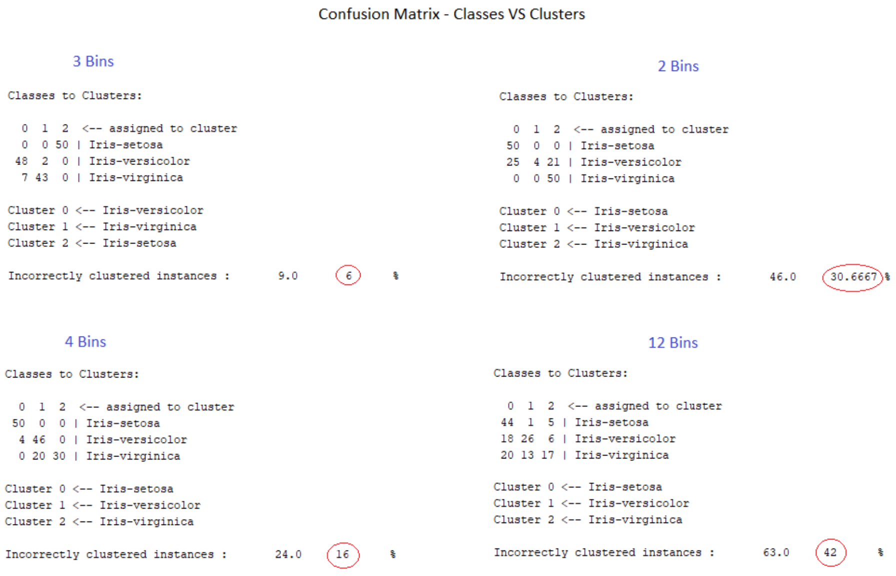
Describing clustering through association analysis

From these we can see that the first cluster is mostly characterized by the dimensions of the petals, while the second and third clusters are characterized by a mix of attribute values. The clustering however is not so good, because 31.3% of the flowers are in the incorrect cluster. It could be that, because we have a higher "resolution" in the possible values that the attributes can take, some sets of attribute values do not reach minimum support and don't give rise to the correct rules. This highlights that it may be important not to overdo the binning process but instead we try to make the discretization in a more significant way, for example by domain knowledge.

Additional Analysis

Different number of bins

For investigating the effect of number of bins on our analysis we used different number of bins for discretizing the values of attributes. The confusion matrix for K-means clustering (Classes to clusters evaluations) changed as bellow:



Confusion Matrix - Classes VS Clusters

3 Bins

```
Classes to Clusters:

  0  1  2  <-- assigned to cluster
  0  0 50 | Iris-setosa
 48  2  0 | Iris-versicolor
  7 43  0 | Iris-virginica

Cluster 0 <-- Iris-versicolor
Cluster 1 <-- Iris-virginica
Cluster 2 <-- Iris-setosa

Incorrectly clustered instances :	9.0	  6	%
```

2 Bins

```
Classes to Clusters:

  0  1  2  <-- assigned to cluster
 50  0  0 | Iris-setosa
 25  4 21 | Iris-versicolor
  0  0 50 | Iris-virginica

Cluster 0 <-- Iris-setosa
Cluster 1 <-- Iris-versicolor
Cluster 2 <-- Iris-virginica

Incorrectly clustered instances :	46.0	 30.6667 %
```

4 Bins

```
Classes to Clusters:

  0  1  2  <-- assigned to cluster
 50  0  0 | Iris-setosa
  4 46  0 | Iris-versicolor
  0 20 30 | Iris-virginica

Cluster 0 <-- Iris-setosa
Cluster 1 <-- Iris-versicolor
Cluster 2 <-- Iris-virginica

Incorrectly clustered instances :	24.0	 16	%
```

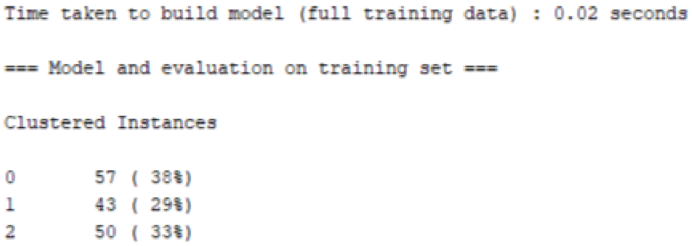
12 Bins

```
Classes to Clusters:

  0  1  2  <-- assigned to cluster
 44  1  5 | Iris-setosa
 18 26  6 | Iris-versicolor
 20 13 17 | Iris-virginica

Cluster 0 <-- Iris-setosa
Cluster 1 <-- Iris-versicolor
Cluster 2 <-- Iris-virginica

Incorrectly clustered instances :	63.0	 42	%
```

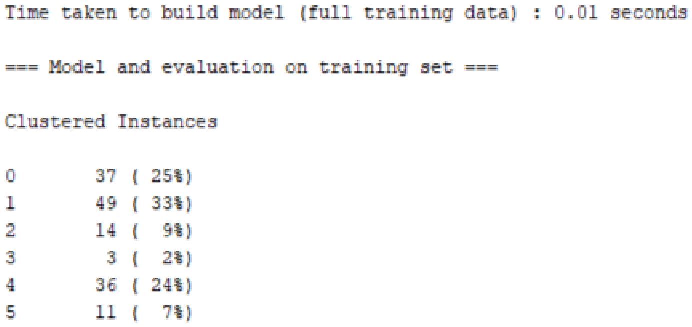
It can be clearly seen when the number if bins increases or decreases the confusion matrix gets worse (The error rate increases) that can result in a worse association analysis.

Different number of clusters

We do not necessarily need to have as many clusters as classes, but if we want to have a reasonable chance of correctly predicting class labels from clusters, we need at least as many clusters as classes. If we have more clusters, it is then not a problem to have more than one cluster map to the same class. These may for instance correspond to further subdivisions of flower types. We experimented with 6 clusters, twice as many as the classes. What happened is that rules satisfying our previously stated constraints were only found for 3 of the clusters: 1, 2 and 5, whereas we do not find any rules for 3, 4 and 6, regardless of the minimum confidence we set.



```
Time taken to build model (full training data) : 0.02 seconds

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

Clustered Instances

0       57 ( 38%)
1       43 ( 29%)
2       50 ( 33%)
```



```
Time taken to build model (full training data) : 0.01 seconds

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

Clustered Instances

0       37 ( 25%)
1       49 ( 33%)
2       14 (  9%)
3        3 (  2%)
4       36 ( 24%)
5       11 (  7%)
```

From this we can conclude that the classes are well separable and that 3 clusters are good enough to obtain an accurate classification. The following rules we found cover each of the clusters. Predictably, if we change the minimum support to 0.01, more of the clusters get rules, with 1, 2, 3, 5, 6 obtaining high confidence rules, as follows.

Table 4: Association rules

Parameters	Cluster	Occurences	Confidience
SL='(5.5-6.7]' PL='(2.96-4.93]'	1	39	1
SL='(5.5-6.7]' PW='(0.9-1.7]'	1	38	1
PL='(4.93-inf)' PW='(1.7-inf)'	2	40	1
SW='(3.6-inf)' PL='(-inf-2.96]'	3	13	1
SW='(2.8-3.6]' PL='(-inf-2.96]'	5	36	1
SW='(2.8-3.6]' PW='(-inf-0.9]'	5	36	1
SL='(-inf-5.5]' SW='(-inf-2.8]' PL='(2.96-4.93]'	6	11	1