Advanced Data Mining Assignment 2: Association Analysis 1

Nazli Bilgic (nazbi056) & Simge Cinar (simci637)

February 22^{th} 2024

1 Introduction

The iris data has 150 rows, and 5 columns. The attributes 'sepallength', 'sepalwidth', 'petallength', and 'petalwidth' are continuous and attribute class is categorical. The scatteplot for the continuous attributes can be seen in the Figure 1. There is a positive correlation between 1) petallength and petalwidth 2) sepallength and petalwidth 3) sepallength and petallength. 'sepalwidth' attribute is not correlated with any other attributes. The graph figures obtained using R and Python for this project.

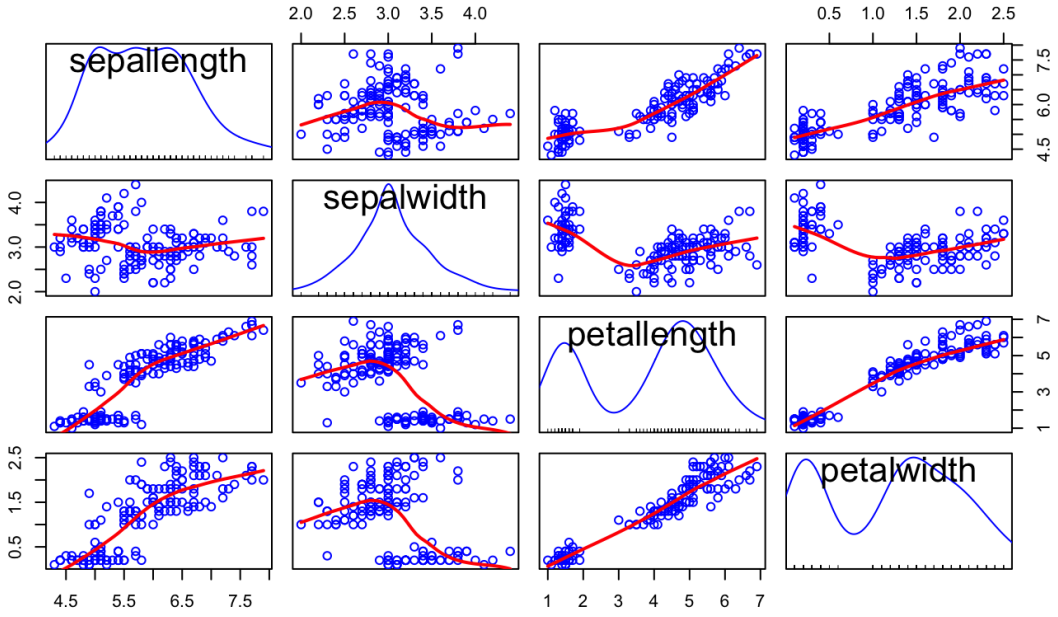

Figure 1: Scatterplot of the continuous attributes

2 Analysis 1

Data is discretized with 3 bins, the histogram for them can be seen in the Figure 2 and 3. Then k-means clustering is applied with 3 clusters for this analysis. Discretized attributes are abbreviated, corresponding variable name for each discretized attribute can be seen in the Table 1.

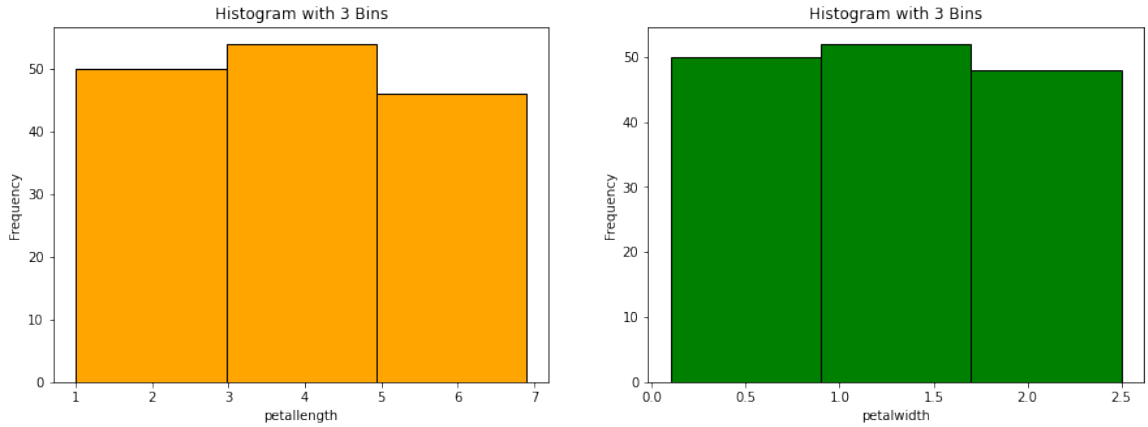

Figure 2: Histogram of petal data with 3 bins

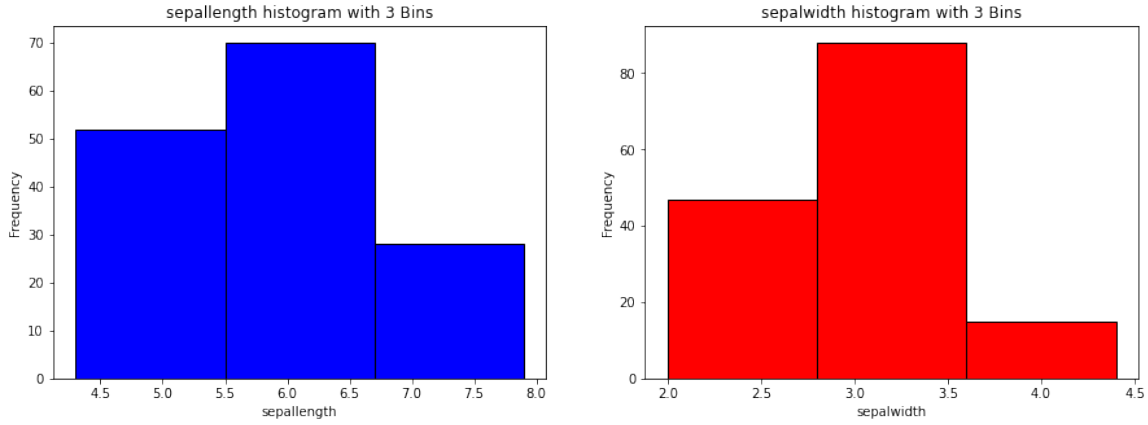

Figure 3: Histogram sepal data with 3 bins

Attribute	Encoding	Count	Attribute	Encoding	Count
petallength='(-inf - 2.966667]'	PL1	50	sepallength="(-inf-5.5]'	SL1	59
petallength='(5.5 - 6.7]'	PL2	54	sepallength='(5.5-6.7]'	SL2	71
petallength='(6.7 - inf)'	PL3	46	sepallength='(6.7-inf)'	SL3	20
petalwidth='(-inf - 0.9]'	PW1	50	sepalwidth='(-inf-2.8]'	SW1	47
petalwidth='(0.9-1.7]'	PW2	54	sepalwidth='(2.8-3.6]'	SW2	88
petalwidth='(1.7 - inf)'	PW3	46	sepalwidth='(3.6-inf)'	SW3	15

Table 1: Encoding for 3 bins

Apriori algorithm is selected for the association analysis. Class attribute is removed and car property is set to True. Class index is -1, which is the created cluster attribute. Metric type is confidence with minimum metric 0.9. lowerBoundMinSupport is 0.1, it means there are no rules whose support is less than 15. Number of rules is selected a large number to get the all rules. 43 rules are generated from the algorithm with given parameters. Rules are sorted first according to their confidence, then their support values. Cluster 1,2 and 3 contains 55, 45 and 50 items respectively. In Weka, the output of algorithm

is indexed and sorted based on the confidence and support value. First, the rules with higher confidence value (generally 1) are displayed. If confidence value is the same, the algorithm sort the rules based on their support value. Some of the rules that are worth to mention can be seen below.

$$\text{Rule 4: } \{\text{PL2, PW2}\}\ (48) \rightarrow \{\text{cluster1}\}\ (48)\ \text{conf:}(1) \tag{1}$$

$$\text{Rule 27: } \{\text{SL2, SW2, PL2, PW2}\}\ (18) \rightarrow \{\text{cluster1}\}\ (18)\ \text{conf:}(1) \tag{2}$$

$$\text{Rule 32: } \{\text{SL2, SW1, PL2, PW2}\}\ (15) \rightarrow \{\text{cluster1}\}\ (15)\ \text{conf:}(1) \tag{3}$$

$$\text{Rule 38: } \{\text{PL2}\}\ (54) \rightarrow \{\text{cluster1}\}\ (51)\ \text{conf:}(0.94) \tag{4}$$

- Rule 4 says that all the items that has PL2 and PW2 are in the cluster 1. It means flowers with petal length between 5.5 and 6.7 and petal width between 0.9 and 1.7 are belong to cluster 1. Support for that rule is 48 and it is the best rule since it is accurate and generalizes well.
- Rule 27 and 32 gives more detail about the cluster. Even though their confidence is high, support values are 18 and 15 which are relatively small. This shows that those rules are not that much general.
- Even though confidence of rule 38 is smaller than the others, its support is high. It can be concluded that this rule is less accurate but more general.
- The attributes of cluster 1 resembles class iris-versicolor. When classes to clusters evaluation is selected as class attribute in the k-means algorithm, 48 items out of 55 in the cluster 1 is assigned to iris-versicolor.

$$\text{Rule 8: } \{\text{PL3, PW3}\}\ (40) \rightarrow \{\text{cluster2}\}\ (40)\ \text{conf:}(1) \tag{5}$$

$$\text{Rule 33: } \{\text{SL2, SW2, PL3, PW3}\}\ (15) \rightarrow \{\text{cluster2}\}\ (15)\ \text{conf:}(1) \tag{6}$$

$$\text{Rule 39: } \{\text{PW3}\}\ (46) \rightarrow \{\text{cluster2}\}\ (43)\ \text{conf:}(0.93) \tag{7}$$

$$\text{Rule 41: } \{\text{PL3}\}\ (46) \rightarrow \{\text{cluster2}\}\ (42)\ \text{conf:}(0.91) \tag{8}$$

- Rule 8 says that all the items that has PL3 and PW3 are in the cluster 2. It means that flowers with petal length between 6.7 and 7.9 and petal width between 1.7 and 2.5 are belong to cluster 2. Support for that rule is 40.
- Rule 33 gives more specific information about 1/3 of the cluster. This rule is accuracte since the confidence is 1 but it is not general, since support is 15
- Confidence value of rule 39 and 41 is lower, they are less accurate but more general rules for cluster 2.

- Rule 8 is the best rule for the cluster 2 with high confidence and support value.
- The attributes of cluster 2 resembles class iris-virginica. When classes to clusters evaluation is selected as class attribute in the k-means algorithm, 43 items out of 45 in the cluster 2 is assigned to iris-virginica.

$$\text{Rule 1: } \{\text{PL1}\}\ (50) \rightarrow \{\text{cluster3}\}\ (50)\ \text{conf:}(1) \tag{9}$$

$$\text{Rule 2: } \{\text{PW1}\}\ (50) \rightarrow \{\text{cluster3}\}\ (50)\ \text{conf:}(1) \tag{10}$$

$$\text{Rule 3: } \{\text{PL1, PW1}\}\ (50) \rightarrow \{\text{cluster3}\}\ (50)\ \text{conf:}(1) \tag{11}$$

$$\text{Rule 7: } \{\text{SL1, PL1}\}\ (47) \rightarrow \{\text{cluster3}\}\ (47)\ \text{conf:}(1) \tag{12}$$

$$\text{Rule 14: } \{\text{SL1, SW2, PL1, PW1}\}\ (36) \rightarrow \{\text{cluster3}\}\ (36)\ \text{conf:}(1) \tag{13}$$

$$\text{Rule 35: } \{\text{SL1, SW2}\}\ (37) \rightarrow \{\text{cluster3}\}\ (36)\ \text{conf:}(0.97) \tag{14}$$

- Rule 3 covers rule 1 and rule 2, support of rule 3 is 0.3
- Rule 14 is more specific, its support is 36. Note that the support value of rules that has 4 elements in other clusters has lower support value. This shows that cluster 3 is more distinguished than others.
- Rule 35 only contains information about sepal attributes, none of the clusters has rules with only sepal since the confidence is not higher than 0.9
- Best rule is 3 with confidence 1 and support value 50.
- The attributes of cluster 3 resembles class iris-setosa. When classes to clusters evaluation is selected as class attribute in the k-means algorithm, 50 items out of 50 in the cluster 3 is assigned to iris-setosa.

Petal attributes which are petallength and petalwidth distiguishes the cluster. Also it should be noticed that best rules for each clusters only contains petal attributes.

3 Analysis 2

Apriori algorithm is selected for the association analysis. Number of clusters changed from 3 to 5 with k-means cluster. Number of bins is 3 and the histogram of the data can be seen in Figure 2 and 3. Discretized attributes are abbreviated, corresponding variable name for each discretized attribute can be seen in the Table 1. Cluster 1, 2, 3, 4 and 5 contains 52, 44, 14, 4 and 36 items. respectively. Class attribute is removed and car property is set to True. Class index is -1, which is the created cluster attribute. Metric type is confidence with minimum metric 0.9. lowerBoundMinSupport is changed from 0.1 to 0.02 since number of cluster is increased and cluster 4 has only 4 items, now there are no rules whose support is less than 3. Number of rules is selected a large number to get the all rules. 69 rules are generated from the

algorithm with given parameters. Rules are sorted first according to their confidence, then their support values.

$$\text{Rule 8: } \{\text{SL2, PL2, PW2}\}\ (33) \rightarrow \{\text{cluster1}\}\ (33)\ \text{conf:}(1) \tag{15}$$

$$\text{Rule 39: } \{\text{SL2, SW1, PL3, PW2}\}\ (4) \rightarrow \{\text{cluster1}\}\ (4)\ \text{conf:}(1) \tag{16}$$

$$\text{Rule 51: } \{\text{SL2, SW1, PL2, PW3}\}\ (3) \rightarrow \{\text{cluster1}\}\ (3)\ \text{conf:}(1) \tag{17}$$

$$\text{Rule 60: } \{\text{PL2, PW2}\}\ (48) \rightarrow \{\text{cluster1}\}\ (45)\ \text{conf:}(0.94) \tag{18}$$

$$\text{Rule 68: } \{\text{PW2}\}\ (54) \rightarrow \{\text{cluster1}\}\ (49)\ \text{conf:}(0.91) \tag{19}$$

- Cluster 1 is the biggest cluster with 52 items, even though rule 8 is good in terms of confidence, it does not generalise well.
- Rule 39 and 51 has really low support value which are 4 and 3 respectively. They give specific information about the cluster but doe not generalise well. Note that the difference between those two rules is the sepal attributes.
- I would say best rule is 60 because it has a high confidence even though it is not 1 and has high support value. This information covers more than 80% of the cluster 1, so as rule 68.

$$\text{Rule 1: } \{\text{PL3, PW3}\}\ (40) \rightarrow \{\text{cluster2}\}\ (40)\ \text{conf:}(1) \tag{20}$$

$$\text{Rule 23: } \{\text{SL2, SW2, PL3, PW3}\}\ (15) \rightarrow \{\text{cluster2}\}\ (15)\ \text{conf:}(1) \tag{21}$$

$$\text{Rule 61: } \{\text{PW3}\}\ (46) \rightarrow \{\text{cluster2}\}\ (43)\ \text{conf:}(0.93) \tag{22}$$

- The best rule for cluster 2 is rule 1 because it is accurate and gives information about 40 items out of 44 items in the cluster 2.
- Rule 23 gives a specific information about approximately 30% of the cluster 2.
- Rule 61 shows that high petal width is a distinguished attribute to cluster 2.

$$\text{Rule 24: } \{\text{SW3, PL1}\}\ (13) \rightarrow \{\text{cluster3}\}\ (13)\ \text{conf:}(1) \tag{23}$$

$$\text{Rule 36: } \{\text{SL1, SW3, PL1, PW1}\}\ (10) \rightarrow \{\text{cluster3}\}\ (10)\ \text{conf:}(1) \tag{24}$$

- Cluster 3 has 14 items, which is lower compared to other clusters.
- No rules with confidence less than 1 is generated for cluster 3.

- Cluster 3 resembles the cluster 5 with higher sepalwidth. Consider the rule 7 below (equation 27), they have the same attributes expect for sepal width but rule 7 is more general than rule 36.

$$\text{Rule 48: } \{\text{SL3, SW2, PW2}\}\ (3) \rightarrow \{\text{cluster4}\}\ (4)\ \text{conf:}(1) \tag{25}$$

- Cluster 4 is the smallest cluster with only 4 elements. It contains the elements with high sepal length.
- Rule 48 is the only generated rule for that cluster.
- This cluster resembles cluster 1 with higher sepal length and width.

$$\text{Rule 2: } \{\text{SW2, PL1}\}\ (36) \rightarrow \{\text{cluster5}\}\ (36)\ \text{conf:}(1) \tag{26}$$

$$\text{Rule 7: } \{\text{SL1, SW2, PL1, PW1}\}\ (36) \rightarrow \{\text{cluster5}\}\ (36)\ \text{conf:}(1) \tag{27}$$

$$\text{Rule 56: } \{\text{SL1, SW2}\}\ (37) \rightarrow \{\text{cluster5}\}\ (36)\ \text{conf:}(0.97) \tag{28}$$

- The all rules above gives information about all items in the cluster 5.
- The best rule is Rule 7, it's accurate with confidence 1 and covers all the elements in the cluster

4 Analysis 3

For this analysis, number of bins is changed from 3 to 5. The histogram for the data can be seen in the Figure 4 and 5. Discretized attributes are abbreviated, corresponding variable name for each discretized attribute can be seen in the Table 2.

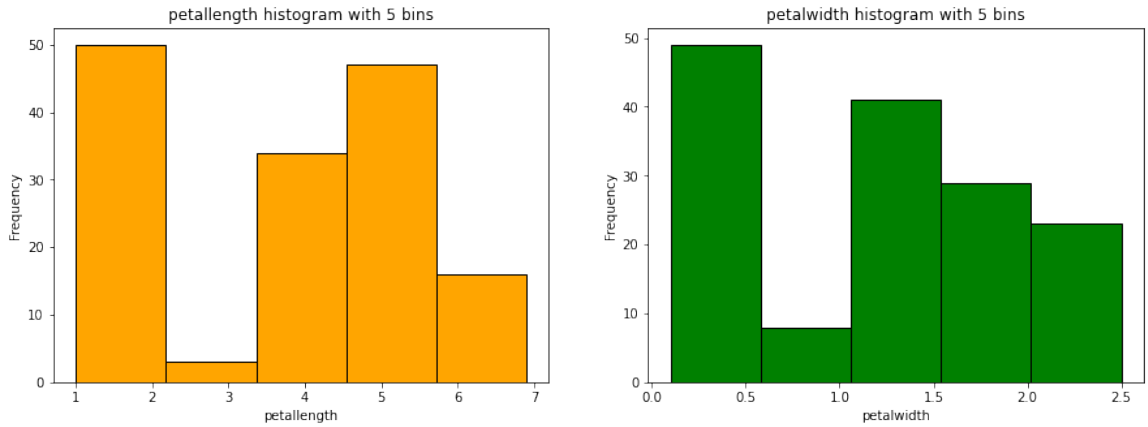

Figure 4: Histogram of petal data with 5 bins

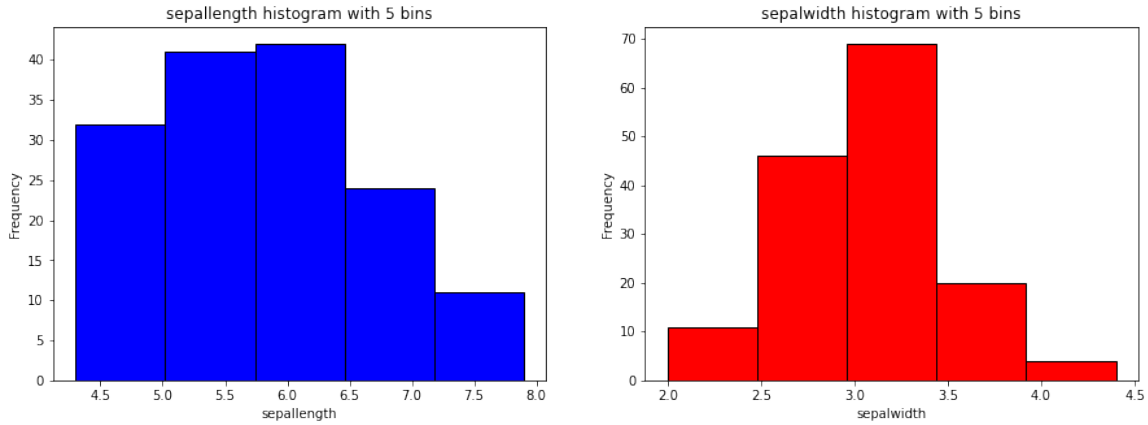

Figure 5: Histogram sepal data with 5 bins

Attribute	Encoding	Count	Attribute	Encoding	Count
petallength='(-inf - 2.18]'	PL1	50	sepallength='(-inf - 5.02]'	SL1	32
petallength='(2.18 - 3.36]'	PL2	3	sepallength='(5.02 - 5.74]'	SL2	41
petallength='(3.36 - 4.54]'	PL3	34	sepallength='(5.74 - 6.46]'	SL3	42
petallength='(4.54 - 5.72]'	PL4	47	sepallength='(6.46 - 7.18]'	SL4	24
petallength='(5.72 -inf)'	PL5	16	sepallength='(7.18 -inf)'	SL5	11
petalwidth='(-inf - 0.58]'	PW1	49	sepalwidth='(-inf-2.48]'	SW1	11
petalwidth='(0.58-1.06]'	PW2	8	sepalwidth='(2.48-2.96]'	SW2	46
petalwidth='(1.06 - 1.54]'	PW3	41	sepalwidth='(2.96-3.44]'	SW3	69
petalwidth='(1.54 - 2.02]'	PW4	29	sepalwidth='(3.44-3.92]'	SW4	20
petalwidth='(2.02 - inf)'	PW5	23	sepalwidth='(3.92-inf)'	SW5	4

Table 2: Encoding for 5 bins

Class attribute is removed and car property is set to True. Class index is -1, which is the created cluster attribute. Metric type is confidence with minimum metric 0.9. lowerBoundMinSupport is 0.1, it means there are no rules whose support is less than 15. Number of rules is selected a large number to get the all rules. 38 rules are generated from the algorithm with given parameters. Then k-means is applied with 3 cluster. Number of elements in each cluster is 63, 35 and 52 respectively.

Rule 6: {SL3, PL4} (27) → {cluster1} (27) conf:(1) (29)

Rule 26: {SL3, SW2, PL4} (16) → {cluster1} (16) conf:(1) (30)

Rule 30: {SL4, SW3, PL4} (15) → {cluster1} (15) conf:(1) (31)

Rule 36: {PL4} (47) → {cluster1} (44) conf:(0.94) (32)

Rule 38: {SL4, SW3} (20) → {cluster1} (18) conf:(0.9) (33)

- Rule 6 is the best rule for cluster 1 in it has a support value of 27.
- From Rule 26 and 30, it can be seen that sepal feature doesn't distinguish the cluster that much. Close sepal length and width values with the same petal length yields to same cluster ???
- Rule 36 is general which has a support value 44.
- Rule 37 has only sepal attributes, there was no rule that contains only sepal feature for cluster virginica in the previous analysis .
- The attributes of cluster 1 resembles class iris-virginica. When classes to clusters evaluation is selected as class attribute in the k-means algorithm, 48 items out of 63 in the cluster 1 is assigned to iris-virginica.

Rule 20: {SW2, PL3} (18) → {cluster2} (18) conf:(1) (34)

Rule 29: {SL2, PL3, PW3} (15) → {cluster2} (15) conf:(1) (35)

Rule 31: {SW2, PL3, PW3} (15) → {cluster2} (15) conf:(1) (36)

Rule 36: {PL3, PW3} (27) → {cluster2} (25) conf:(0.93) (37)

- Rule 20 is good in terms of accuracy but Rule 36 generalizes more.
- The attributes of cluster 2 resembles class iris-versicolor. When classes to clusters evaluation is selected as class attribute in the k-means algorithm, 33 items out of 35 in the cluster 2 is assigned to iris-versicolor.

Rule 1: {PL1} (50) → {cluster3} (50) conf:(1) (38)

Rule 2: {PW1} (49) → {cluster3} (49) conf:(1) (39)

Rule 3: {PL1, PW1} (49) → {cluster3} (49) conf:(1) (40)

Rule 12: {SL1, SW3} (22) → {cluster3} (22) conf:(1) (41)

Rule 15: {SL1, SW3, PL1, PW1} (22) → {cluster3} (22) conf:(1) (42)

Rule 34: {SL1} (32) → {cluster3} (30) conf:(0.94) (43)

- Rule 1,2 and 3 generalizes well, the best rule is 3.
- Cluster 3 is more distinguished than others because its support values are high with high confidence levels.
- The attributes of cluster 3 resembles class iris-setosa. When classes to clusters evaluation is selected as class attribute in the k-means algorithm, 50 items out of 52 in the cluster 3 is assigned to iris-setosa.

5 Conclusion

- In the introduction part, we mentioned that there is positive correlation between some attributes. It was observed that correlated variables are grouped together in the ancedent set. For example, we did not observe a set like {PL1, PW5}, small petal length is not together with small petal width in any sets because there is positive correlation between attribute petallength and petalwidth. The flowers with high petallength also has high petalwidth.

- Comparing analysis 1 and 2, we increased the number of cluster from 3 to 5. To get rules for the small cluster, like cluster 4, we decreased minimum support from 0.1 to 0.02 and we get less general rules. Cluster 4 resembles cluster 1 with higher sepallength and sepalwidth. When we look at the histograms for sepal attribute in Figure 3, it can be observed that number of elements with high sepallength and sepalwidth is small. Moreover, we know that we originally have 3 classes hence this cluster might contains the outlier items in terms of sepal attribute.

- Comparing analysis 1 and 3, increasing number of bins resulted in less rules whatsoever I would expect otherwise. This shows that when we increased the number of rules, some rules didn't exceed the minimum support or confidence. The support values in the analysis with higher bin is lower which means increasing number of bins might cause worse generalization.

- When number of bins is 3, 33 rules have confidence 1 out of 43 rules (76.74%) and when number of bins is 5, 31 rules have confidence 1 out of 38 rules (81.58%). Increasing number of bins resulted in less rules but generated rules which are more accurate (with the same min support and min confidence). Also it can be observed that increasing number of bins caused lower support values in the best rules.

- The visualization of the attributes petallength and petalwidth for analysis 1 (3 cluster with 3 bins) using Weka can be seen in the Figure 6. As mentioned earlier, there is a positive correlation between attributes petallength and petalwidth. During the analysis 1, it is demonstrated that petal attributes differ in each cluster. From the graph, it can be observed that flowers with small petallength has small petalwidth and they are clustered together, same logic applies to middle and large petal values as well.

- The visualization of the attributes sepallength and sepalwidth for analysis 1 (3 cluster with 3 bins) using Weka can be seen in the Figure 7. It can be observed that sepalwidth does not differentiate cluster 3 (green) much, both middle and high sepalwidth values result in cluster 3. Moreover, it can be seen that cluster 1 (blue) and cluster 2 (red) is not differentiated by the sepal attributes. Cluster 1 has low middle and high sepallength and sepalwidth values, so as cluster 2.

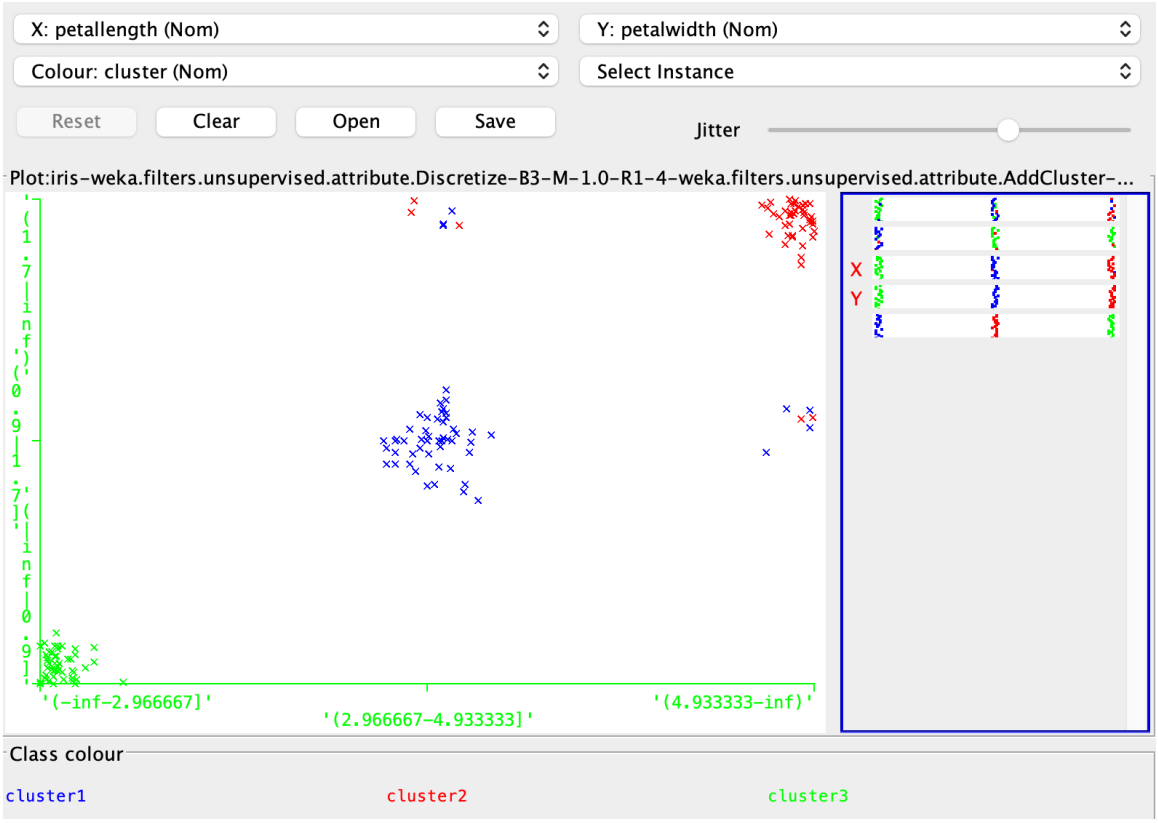

Figure 6: Visualization of discretized petallength and petalwidth

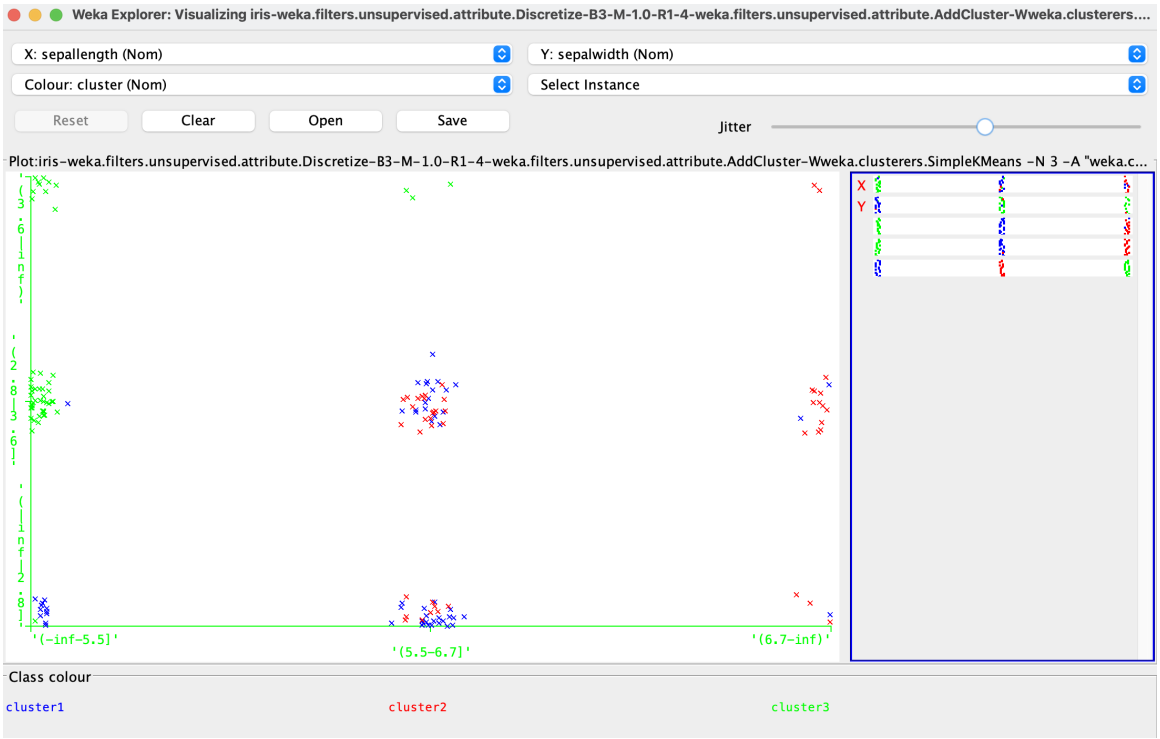

Figure 7: Visualization of discretized sepallength and sepalwidth