Comparing results and performance: association rule analysis and classification

**Include results/answers for the items in red.**

*Comparing the results of different data mining approaches. A sample dataset is posted on Canvas (zoo dataset).* You can use either R or Weka (include code and any results / visualizations).

(A) Analyze the dataset with decision trees (i.e. J48 in Weka, or ctree in R)

**[What percentage of instances are correctly classified by your decision tree?]**

**92.9293 %**

**[Which classes (in the zoo case, families) are often mistaken for each other?]**

**Reptiles and amphibians**

(B) Experiment with some of the other classifiers (Naïve Bayes, Random Forest) and see if we can get a better classification performance.

**[Which classifier had the best performance and why? Which performance measures support your claim?]**

**Naïve Bayes: 93.9394 %**

**Random Forest: 90.9091 %**

**Decision Table: 84.8485 %**

**Naïve Bayes had the best performance because it correctly classified the classes with the most accuracy. The results for each of the other classifiers all support the claim that reptiles and amphibians are mistaken for each other based on the confusion matrix. In all cases, there seems to be an overlap between the two.**

**[Is any preprocessing needed before running the Apriori algorithm? Hint- for any numerical attributes.]**

**Yes, we need to change the legs column to be a boolean type. If the number of legs is 0 then the column should be false. If the number of legs is greater than 0, then the column value should be true.**

(C) Analyze with Apriori

**[Which analysis was most/least useful for the dataset, and why? Include relevant results.]**

**Output from Weka:**

Apriori

=======

Minimum support: 0.7 (69 instances)

Minimum metric <confidence>: 0.9

Number of cycles performed: 6

Generated sets of large itemsets:

Size of set of large itemsets L(1): 9

Size of set of large itemsets L(2): 16

Size of set of large itemsets L(3): 4

Best rules found:

1. venomous=False tail=True 71 ==> backbone=True 71 <conf:(1)> lift:(1.22) lev:(0.13) [12] conv:(12.91)

2. legs=True 76 ==> fins=False 75 <conf:(0.99)> lift:(1.19) lev:(0.12) [12] conv:(6.53)

3. tail=True 75 ==> backbone=True 74 <conf:(0.99)> lift:(1.21) lev:(0.13) [12] conv:(6.82)

4. venomous=False legs=True 72 ==> fins=False 71 <conf:(0.99)> lift:(1.19) lev:(0.11) [11] conv:(6.18)

5. breathes=True legs=True 71 ==> fins=False 70 <conf:(0.99)> lift:(1.19) lev:(0.11) [11] conv:(6.1)

6. backbone=True tail=True 74 ==> venomous=False 71 <conf:(0.96)> lift:(1.04) lev:(0.03) [2]

conv:(1.49)

7. backbone=True 81 ==> venomous=False 77 <conf:(0.95)> lift:(1.03) lev:(0.03) [2] conv:(1.31)

8. breathes=True 78 ==> fins=False 74 <conf:(0.95)> lift:(1.15) lev:(0.09) [9] conv:(2.68)

9. legs=True 76 ==> venomous=False 72 <conf:(0.95)> lift:(1.03) lev:(0.02) [2] conv:(1.23)

10. airborne=False 75 ==> feathers=False 71 <conf:(0.95)> lift:(1.19) lev:(0.11) [11] conv:(3.03)

Answer: I believe that the Naïve Bayes classifier was the most useful one because it provided the best confidence value when determining the zoo class.