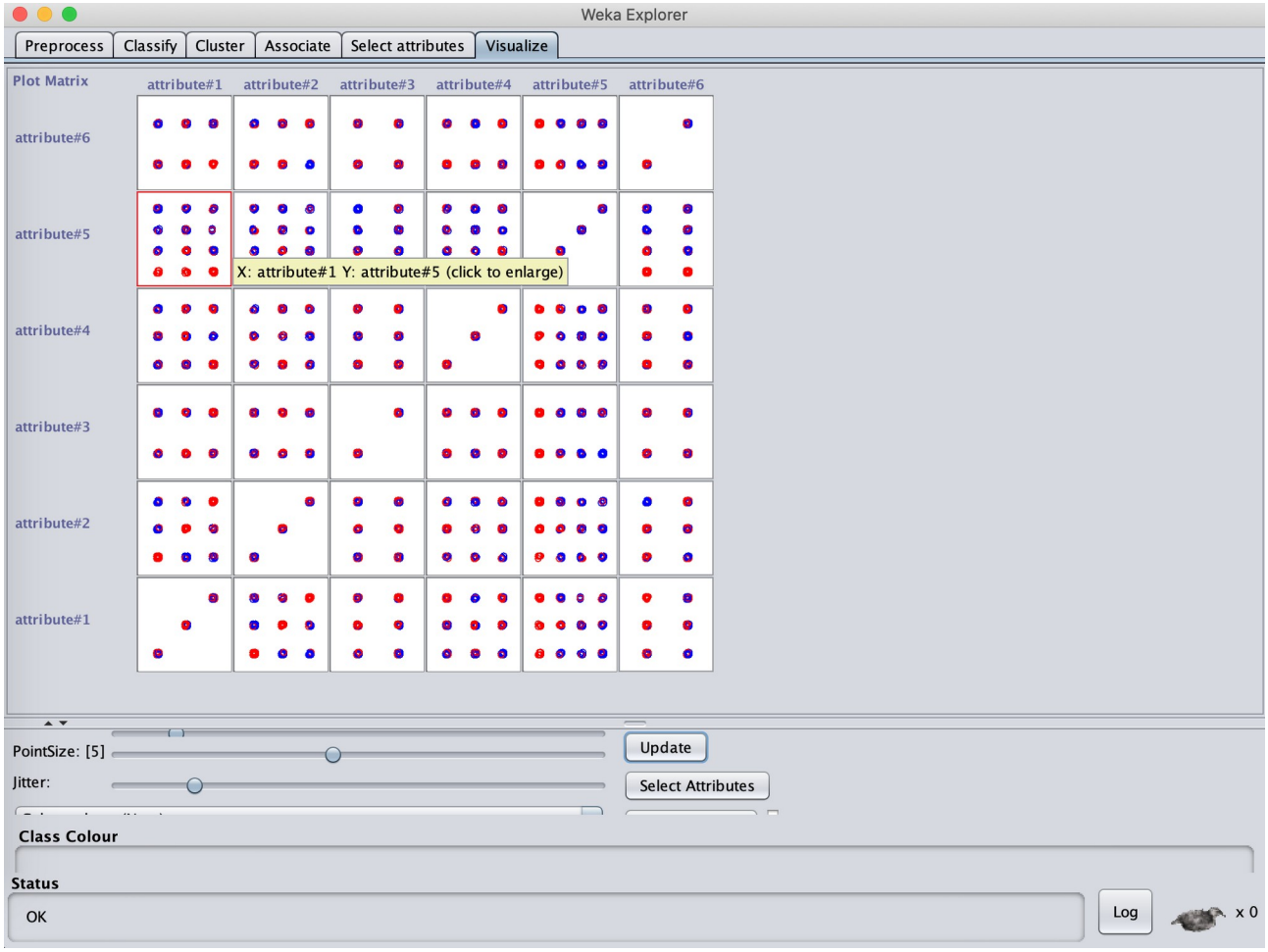
Data Mining
Lab3 - Association Analysis - 2

Zhixuan_Duan(zhidu838)

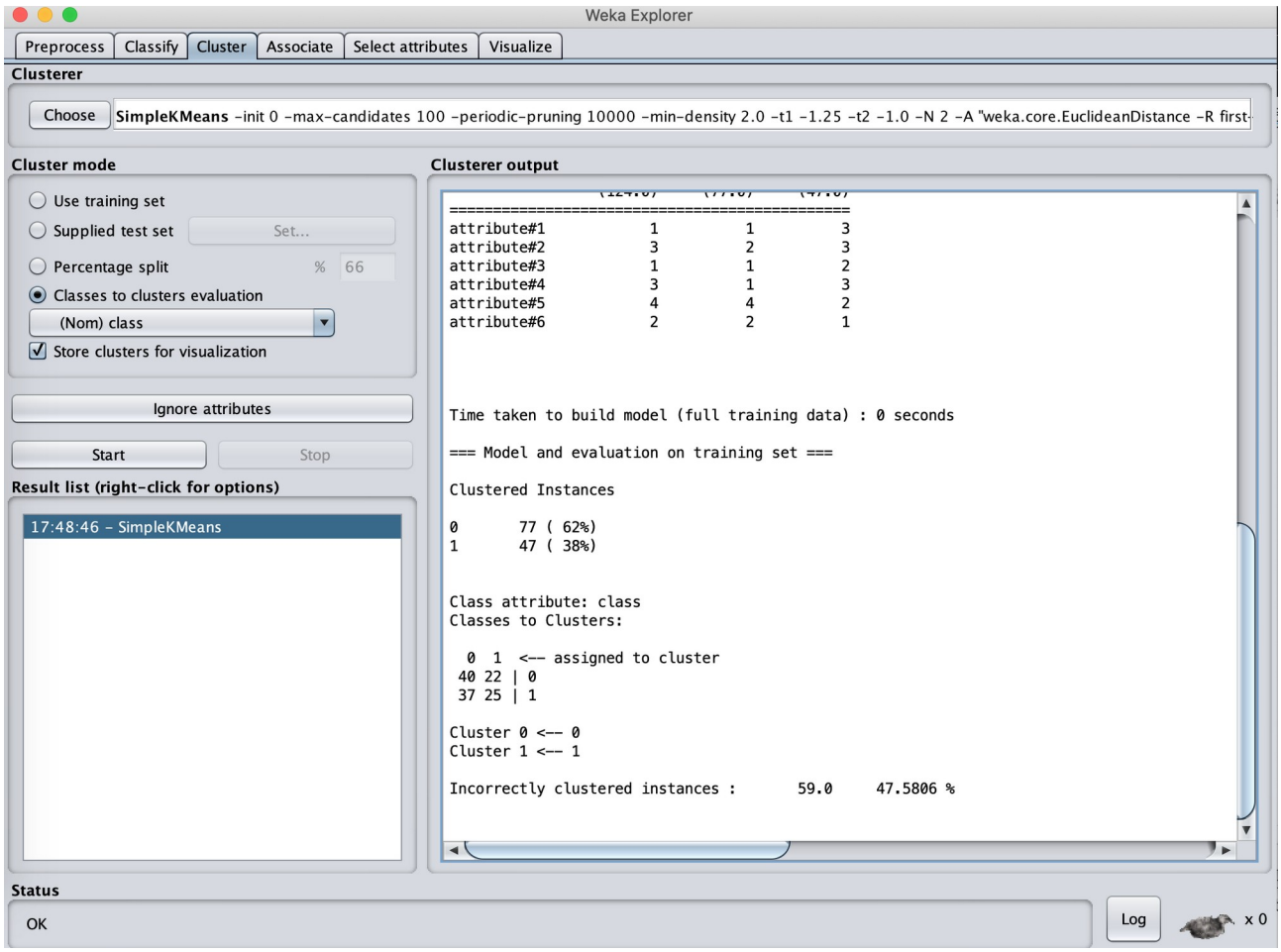
2020-03-01

1. Clustering

Before clustering, we plot the dataset by visualization, and it is pretty clear there are lot of overlap, and lots of similar attributes share different class.



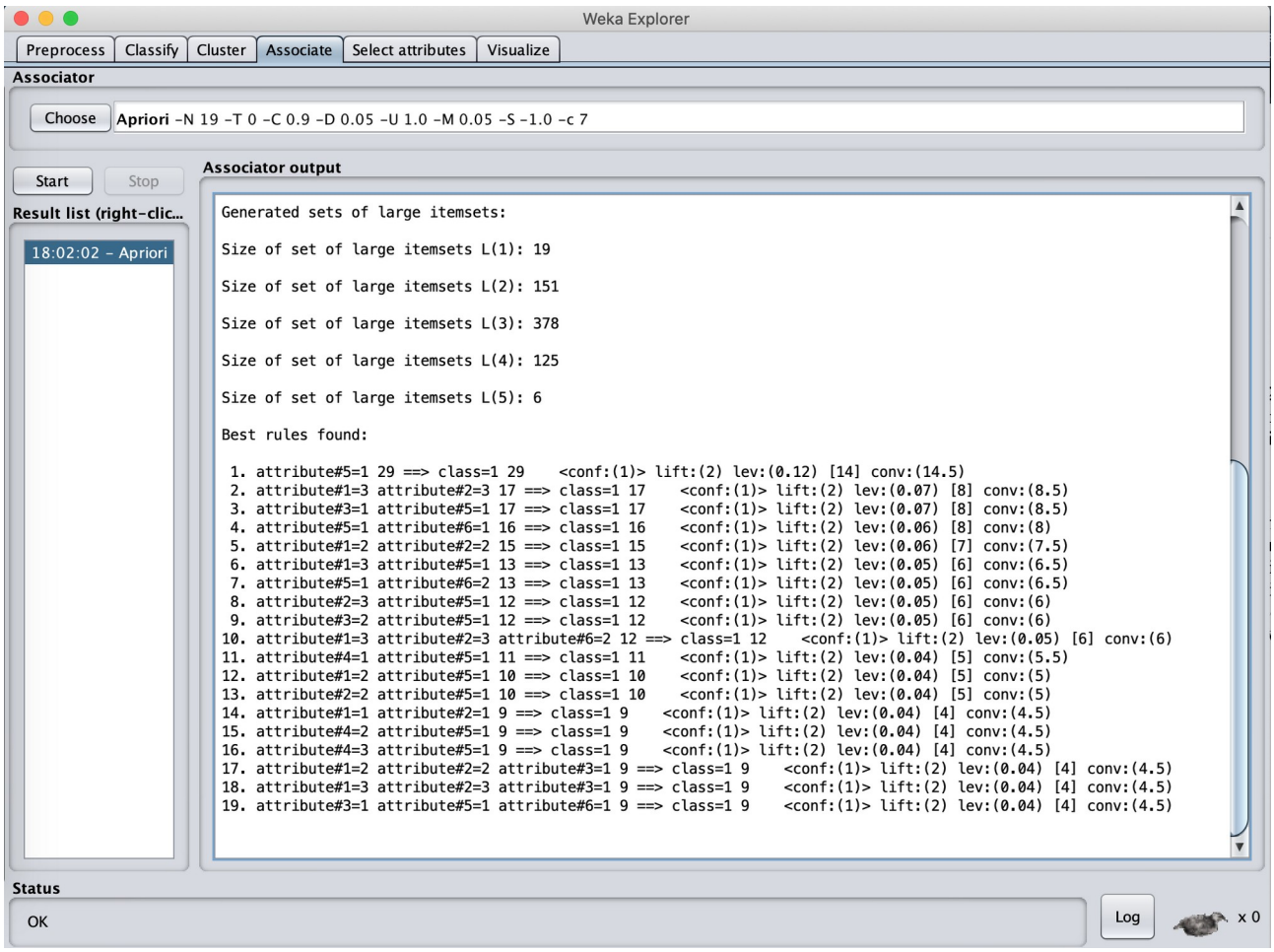
And we choose K-means clustering method, here is the results.



Analysis: From the output, we get 47% incorrectly clustered instances, indicates a bad clustering.

2. Association Analysis

Then we perform association analysis, here is the results.



After remove remove redundant rules, we got the following four rules:
1. attribute#5=1 29 ==> class=1 29 <conf:(1)> lift:(2) lev:(0.12) [14] conv:(14.5)
2. attribute#1=3 attribute#2=3 17 ==> class=1 17 <conf:(1)> lift:(2) lev:(0.07) [8] conv:(8.5)
3. attribute#1=2 attribute#2=2 15 ==> class=1 15 <conf:(1)> lift:(2) lev:(0.06) [7] conv:(7.5)
4. attribute#1=1 attribute#2=1 9 ==> class=1 9 <conf:(1)> lift:(2) lev:(0.04) [4] conv:(4.5)

3. Why can the clustering algorithms not find a clustering that matches the class division in the database?

Analysis:
Clustering algorithms need to choose the initial node and define the distance formula to judge the similarity, in K-means method, both E and M distance can not work for this dataset since there are a lot of overlap data, and under this situation, clustering can not find an optimal clustering to matches the class division in the database.

4. Would you say that the clustering algorithms fail or perform poorly for the monk1 dataset? Why or why not?

Analysis:
Under traditional methods like k-means, it won't work, however, if add one new attribute computed by kernel method, try to convert low-dimensional inseparable to high-dimensional linearly separable data, then use distance to judge the similarity, it may work.