Methods of data mining

Assignment 3

Task 2

Miro-Markus Nikula

712686

a) I used the existing binarization for ratdata and binarized the following features: sulcer, gonind, appind and heartind. For kmethod I couldn't really find the meanings of the numeral codes. In one description it said that 1-5 the kmethod is unknown, but those were the only values displayed in the data for this exercise and for that reason, I chose not to use kmethod. The new features derived from the aforementioned features are gonindlow(<0.05), gonindhigh(>1.5), appindlow(<0.0099), appindhigh(>0.0177), heartindlow(<0.0035), heartindhigh(>0.0048), hassulcer(>1). For the index-features I took the 20^{th} and 80^{th} percentiles using Numpy to determine which values correspond to high- and low-features. I used basic Pandas' dataframe methods together with python's text-file and string-methods.

b)

I transferred the transaction data into numerical codes using namescodes-tool and then found 100 positive association rules using command: ./kingfisher -i test.txt.codes -k599 -M-50 -t1 -o testrules.txt.

The 11 rules I took from the top 100 are:

1. **place2 -> winter** fr=62 (0.1105), cf=0.827, gamma=3.336, delta=0.077, M=-6.892e+01
2. **wild hassulcer -> heartindhigh** fr=51 (0.0909), cf=0.354, gamma=3.896, delta=0.068, M=-7.713e+01
3. **wild adrenallarge -> hassulcer** fr=72 (0.1283), cf=0.713, gamma=2.666, delta=0.080, M=-5.788e+01
4. **heartindhigh -> hassulcer** fr=51 (0.0909), cf=1.000, gamma=3.740, delta=0.067, M=-7.459e+01
5. **freezer -> heartindlow** fr=31 (0.0553), cf=0.912, gamma=5.384, delta=0.045, M=-5.141e+01
6. **freezer -> liversmall** fr=29 (0.0517), cf=0.853, gamma=6.214, delta=0.043, M=-5.086e+01
7. **lab -> gonindhigh** fr=43 (0.0766), cf=0.597, gamma=3.527, delta=0.055, M=-4.296e+01
8. **liversmall lab -> gonindhigh** fr=30 (0.0535), cf=0.833, gamma=4.921, delta=0.043, M=-4.391e+01
9. **weightlow -> heartindhigh** fr=61 (0.1087), cf=0.616, gamma=3.032, delta=0.073, M=-5.477e+01
10. **summer weightnormal -> mother-female** fr=168 (0.2995), cf=0.530, gamma=1.509, delta=0.101, M=-5.767e+01
11. **summer weightnormal place3 -> nursing** fr=96 (0.1711), cf=0.430, gamma=2.047, delta=0.088, M=-5.687e+01

c) I chose these rules mainly from biological point of view and preferred the ones that I made the most sense. Place2 seems to be a common one for finding rats during winter (rule1). Wild and sulcer seem to appear together quite often (rule2-4). One hypothesis could be that wild rats are more likely to have sulcer and sulcer is connected to higher heartind. Doesn't sound too far off to me since stomach ulcer makes digesting food difficult, which then leads to body weight loss.

Freezer rats seem to have smaller hearts and livers (rule5-6). Also, some lab rats have higher gonind and about 2/3 of them also have a small liver (rule7-8). One very obvious association is in rule 9. If body weight is low, then most likely heartind(=heart weight / body weight) is high.

The most frequent associations here are the last two. Summer is a time when mother rats and nursing rats can be found. In many cases also from the same place. Without diving any deeper into this matter, I believe summer is the time when most animals, especially rodents breed. The nursing period of a rat is about one month which would explain why they're also captured during summer. Normal weight is also associated here and that I believe is because the mothers are no longer pregnant.