Exploring Caterpillar dataset

Amongst numerous freely available datasets in R I found a Caterpillar dataset (from library Stat2Data) that contains the measurements of the physical state of caterpillar such as body mass, food intake and nitrogen assimilation. There are also some categorical attributes that describe the life stage of the animal. A dataset has 267 observations for 18 variables.

First thing that we want to do is to get familiar with the observations. As the dataset contains multiple features that we are not able to plot as N-dimensional plot, one option is to create a multiple 2D plots and compare them.

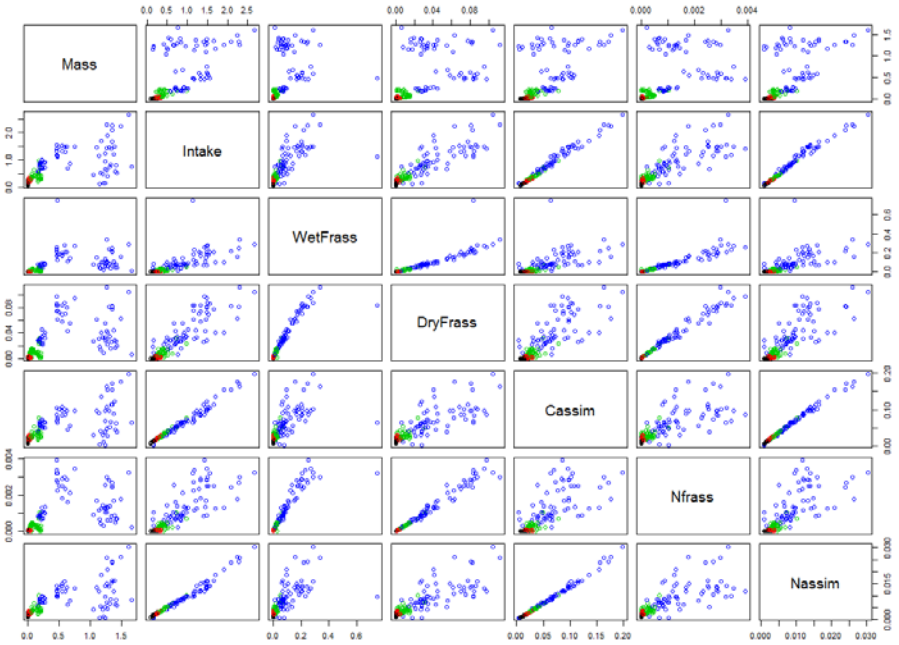

Abbildung 1 Scatter plot of all non-categorical attributes in our dataset. Data points are limited to only the first 4 stages to highlight the similarities of measurements in those stages.

Dimensionality reduction

With this approach it is hard to get the grasp of data clumpiness. Another approach is to reduce the dimensionality of the dataset to 2D or 3D as we can easily visualize this. For this I used transformation method called tSNE that is one of the best and widely used dimensionality reduction technique.

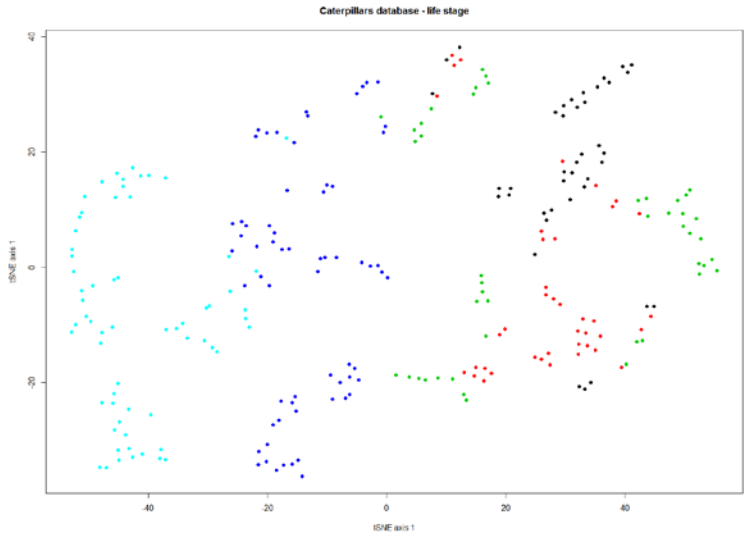

Abbildung 2 Results of tsne projection colored by the stage of caterpillar's life.

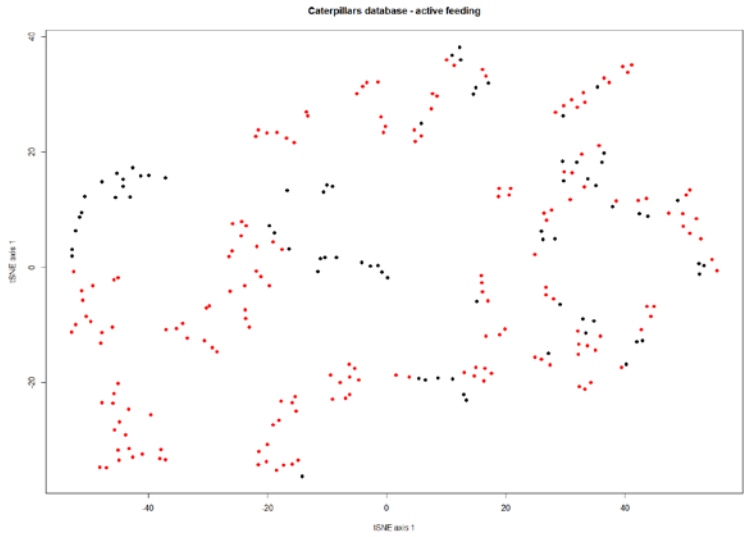

Abbildung 3 Results of tsne projection colored by the status of active feeding.

As we can see from the previous two scatter plots of the projection, main driving attribute behind the measurements is the stage of caterpillar's life. We can see there are multiple clusters of measurement that can be attributed to stage attribute. The next best attribute that could describe further clumpiness of data is the stage of active feeding.

Classification – active feeding

I was also interested in how good we can classify the feeding process using the physical measurements of the animals. For this I used the Decision tree classifier from party library. The accuracy of the classification was 84%.

Contingency table of the classification procedure:

```
     pred
train   1   2
    1  42  32
    2   8 171
```

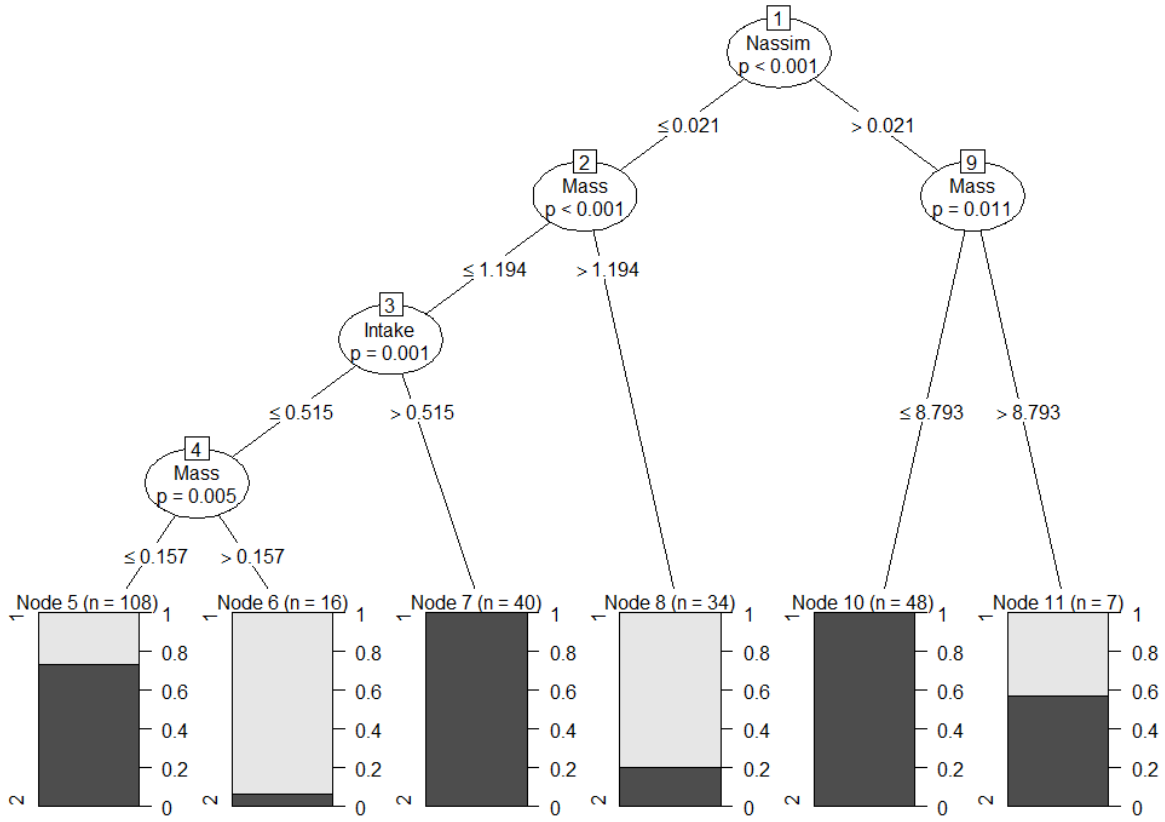

Abbildung 4 Visualization of decision tree.