Can machine Learning be used to identify species of Sorbus

PetraGuy

January 20 2018

```
#This creates pdf from command line, note, sensitive to ' or "

#Rscript -e "library(knitr); knit('MiniProj2.Rmd')"

#Rscript -e "library(rmarkdown); render('MiniProj2.md')"
```

Within cluster sum of squares/between cluster sum of squares for the unscaled, semi-scaled and fully scaled data for ten repeats of kmeans.

	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10
unscaled	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14
semi-scaled	0.34	0.34	0.34	0.34	0.34	0.34	0.34	0.34	0.34	0.34
fullyscaled	0.42	0.42	0.42	0.42	0.42	0.42	0.42	0.42	0.42	0.42

The ratio is largest for the scaled data and does not decrease when the data is scaled, so an unscaled data set is preferable. The ratio is identical each time.

The accuracy over the ten repeats.

	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10
unscaled	0.18	0.10	0.08	0.20	0.05	0.17	0.10	0.09	0.07	0.06
semi-scaled	0.12	0.13	0.03	0.12	0.07	0.07	0.20	0.10	0.12	0.22
fullyscaled	0.08	0.11	0.23	0.11	0.02	0.16	0.17	0.26	0.05	0.06

The accuracy is different for each repeat on all the data sets, suggesting that the algorithm is not successfully clustering the data.

The next tables show the percentage of each species correctly allocated to its cluster on each of the ten repeats. For example, the top row from left to right, gives the true positive rate for S. anglia on each subsequent run on the kmeans algorithm.

unscaled

	1	2	3	4	5	6	7	8	9	10
Anglica	31.87	0.00	0.00	31.87	0.00	13.12	13.12	0.00	0.00	0.00
Cuneifolia	0.00	2.00	22.00	34.00	6.00	6.00	2.00	4.00	4.00	4.00
Intermedia	0.00	0.00	63.16	0.00	0.00	5.26	0.00	26.32	0.00	0.00
Leyana	29.17	20.83	8.33	2.08	29.17	31.25	6.25	2.08	31.25	29.17
Minima	3.33	6.67	16.67	3.33	6.67	3.33	3.33	6.67	36.67	6.67
Mougeotii	8.00	50.00	0.00	8.00	0.00	46.00	4.00	50.00	0.00	8.00
Arranensis	0.00	4.35	0.00	4.35	0.00	4.35	43.48	0.00	0.00	0.00

semi-scaled

	1	2	3	4	5	6	7	8	9	10
Anglica	10.62	29.38	1.25	5.62	1.25	1.25	9.38	0.62	5.62	30.63
Cuneifolia	8.00	0.00	10.00	2.00	0.00	30.00	8.00	22.00	22.00	0.00
Intermedia	47.37	0.00	0.00	0.00	0.00	0.00	52.63	52.63	47.37	0.00
Leyana	2.08	2.08	2.08	20.83	47.92	18.75	0.00	18.75	0.00	20.83
Minima	0.00	0.00	0.00	83.33	0.00	0.00	83.33	23.33	10.00	0.00
Mougeotii	30.00	2.00	0.00	0.00	4.00	0.00	4.00	2.00	30.00	46.00

```
Arranensis  0.00  0.00  8.70  0.00  0.00  0.00 91.30  0.00  0.00  0.00
```

```
scaled
```

```
                1      2     3      4    5     6     7     8  9    10
Anglica     0.0   0.00 26.88   0.00 0.00 18.12  0.00 31.25  0  0.00
Cuneifolia  2.0   2.00 48.00   6.00 0.00  6.00 52.00 52.00 34  6.00
Intermedia  0.0   0.00 94.74   0.00 0.00  0.00  5.26 94.74  0 94.74
Leyana     62.5   2.08  4.17  20.83 4.17 12.50 20.83  6.25  0  4.17
Minima      0.0   0.00  3.33   0.00 6.67  0.00 83.33  0.00  0  0.00
Mougeotii   0.0  32.00  0.00  10.00 6.00  0.00  8.00  0.00  6  0.00
Arranensis  0.0 100.00  0.00 100.00 0.00 91.30  0.00  0.00  0  0.00
```

The results again show that the algorithm is not consistently allocating species to the correct cluster. On some runs, is is very accurate for some species, but not necessarily for all the others. Then on other runs its is completely inaccurate.

Hierarchical Clustering.

First the unscaled imputed data.

```
Accuracy obtained using the different distance calculations
```

```

Distance Method "euclidean" "maximum" "manhattan" "canberra" "minkowski"
Accuracy        "0.3"       "0.18"    "0.29"      "0.42"     "0.3"
```

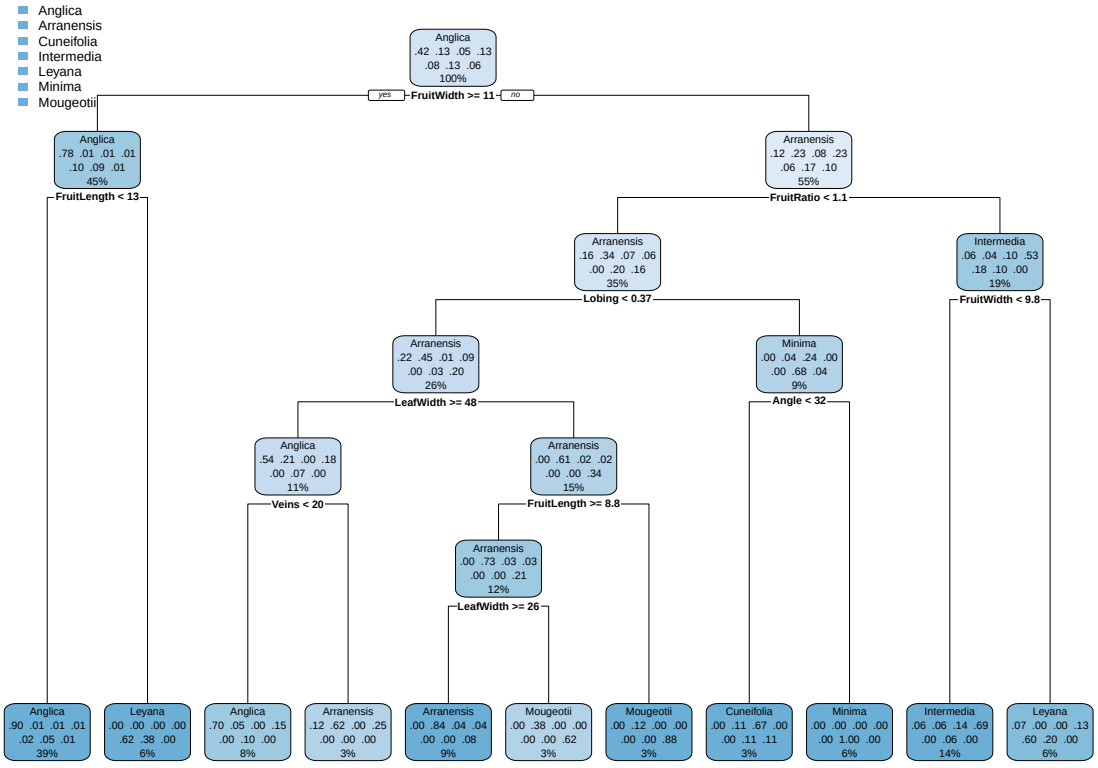
The canberra distance gives the most accurate results, below is the confusion matrix for that method

```
             cluster
                1   2   3   4   5   6   7
  Anglica     108  41  11   0   0   0   0
  Arranensis    7  14   0   0  22   0   7
  Cuneifolia    0   1   0   8  10   0   0
  Intermedia   13   2   2  22   8   1   0
  Leyana        4  17   6   0   0   3   0
  Minima        1   8   9  29   0   3   0
  Mougeotii     0   0   0   0  12   0  11
```

The confusion matrix shows that whilst many S anglica and S minima were clustered together, many other species were dispersed across clusters. The model is not efficient in separting the species.

When the scaled date was used, the accuracy reduced. The results demonstrate that while hierarchical clustering was more accurate than kmeans when the canberra distance is used, the model is still very poor.

Decision Tree



The decision tree produces some very accurate nodes. 100 % for S. minima, 90% for S. anglica,

The accuracy for the decision tree is

0.68

The sensitivies for the decision tree model are shown in the table below

```
     Species      Sensitivity
[1,] "Anglica"    "0.85"
[2,] "Arranensis" "0.47"
[3,] "Cuneifolia" "0.17"
[4,] "Intermedia" "0.71"
[5,] "Leyana"     "0.67"
[6,] "Minima"     "0.6"
[7,] "Mougeotii"  "0.43"
```

The precisions for the decision tree model are shown in the table below

```
     Class              Precision
[1,] "class_Anglica"    "0.79"
[2,] "class_Arranensis" "0.5"
[3,] "class_Cuneifolia" "0.5"
[4,] "class_Intermedia" "0.67"
[5,] "class_Leyana"     "0.4"
[6,] "class_Minima"     "1"
[7,] "class_Mougeotii"  "0.43"
```

The confusion matrix details exactly how the species were placed.

The final column is the sum of species in the test set.

```
           Anglica Arranensis Cuneifolia Intermedia Leyana Minima
Anglica         41          3          0          0      3      0
```

```
Arranensis       4          7          0          0      1      0
Cuneifolia       0          1          1          4      0      0
Intermedia       2          0          1         10      1      0
Leyana           3          0          0          0      6      0
Minima           1          0          0          1      4      9
Mougeotii        1          3          0          0      0      0
            Mougeotii
Anglica           1 48
Arranensis        3 15
Cuneifolia        0  6
Intermedia        0 14
Leyana            0  9
Minima            0 15
Mougeotii         3  7
```