Can machine Learning be used to identify species of Sorbus

PetraGuy

January 18 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.

```
            [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
unscaled    0.14 0.14 0.14 0.14 0.14 0.14 0.14 0.14 0.14  0.14
semi-scaled 0.35 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.

```
            [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
unscaled    0.21 0.17 0.11 0.14 0.21 0.17 0.13 0.14 0.18  0.10
semi-scaled 0.15 0.23 0.21 0.14 0.11 0.14 0.17 0.09 0.06  0.09
fullyscaled 0.20 0.13 0.10 0.22 0.18 0.08 0.15 0.07 0.03  0.10
```

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    31.25 17.50  3.12 23.75 28.75 13.12 11.88 11.88 12.50  3.12
Cuneifolia 20.00 20.00  6.00  4.00  6.00 40.00 34.00  6.00  4.00 30.00
Intermedia  5.26  0.00 21.05 57.89 21.05  0.00  5.26 52.63 26.32  0.00
Leyana      2.08 10.42 18.75  4.17 18.75 18.75  6.25  2.08 31.25 29.17
Minima      3.33 23.33  3.33  3.33 36.67  3.33 23.33  3.33  3.33  3.33
Mougeotii  16.00  6.00  4.00  0.00  0.00 28.00  4.00  8.00 50.00  4.00
Arranensis 43.48 52.17 73.91  4.35 21.74  0.00  0.00 73.91  0.00  0.00
```

semi-scaled

```
               1     2     3     4     5     6  7     8     9    10
Anglica    11.25 30.63 21.25  3.12  1.25 11.25 30 11.25  1.25 19.38
Cuneifolia 26.00  8.00 22.00 22.00 30.00  0.00 26  0.00 26.00  0.00
```

Intermedia	0.00	0.00	52.63	26.32	0.00	52.63	0	0.00	0.00	0.00
Leyana	0.00	20.83	20.83	31.25	2.08	20.83	0	2.08	18.75	2.08
Minima	83.33	0.00	3.33	3.33	3.33	3.33	0	0.00	0.00	0.00
Mougeotii	2.00	46.00	30.00	30.00	4.00	28.00	4	30.00	0.00	2.00
Arranensis	0.00	0.00	0.00	0.00	91.30	0.00	0	8.70	0.00	0.00

scaled

	1	2	3	4	5	6	7	8	9	10
Anglica	28.12	0.00	0.00	26.88	27.50	0.00	28.75	0.00	1.88	1.88
Cuneifolia	10.00	2.00	16.00	4.00	4.00	4.00	8.00	46.00	0.00	50.00
Intermedia	0.00	0.00	0.00	94.74	100.00	0.00	0.00	0.00	0.00	0.00
Leyana	8.33	2.08	6.25	4.17	4.17	4.17	4.17	4.17	6.25	12.50
Minima	0.00	83.33	6.67	3.33	3.33	83.33	6.67	0.00	6.67	6.67
Mougeotii	0.00	48.00	50.00	32.00	0.00	6.00	4.00	6.00	6.00	6.00
Arranensis	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	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.