Cluster Analysis: Identifying Parkinson's Disease Subtypes

July 22, 2015

1 Preprocessing

1.1 Dataset Description

951 subjects, 145 metrics, collected 15-4-2012 from Pablo Martinez Martín. Only 19 features used for clustering and/or interpretation. 50 subjects with missing values of the features to be used in clustering (brought down to 901). Imputation may be a good idea later on.

1.2 Selected Features

Combination of non-motor scale (NMS) symptoms and standard motor symptoms. Note: PIGD was deleted after 2015-07-16 meeting.

Name	Type	Format	Description
nms_d1	byte	%8.0g	cardiovascular
nms_d2	byte	%8.0g	sleep/fatigue
nms_d3	byte	%8.0g	mood/cognition
nms_d4	byte	%8.0g	percep/hallucinations
nms_d5	byte	%8.0g	attention/memory
nms_d6	byte	%8.0g	gastrointestinal
nms_d7	byte	%8.0g	urinary
nms_d8	byte	%8.0g	sexual function
nms_d9	byte	%8.0g	miscellaneous
tremor	float	%9.0g	tremor
bradykin	float	%9.0g	bradykinesia[1]
rigidity	float	%9.0g	rigidity
axial	float	%9.0g	axial[2]

Table 1: Selected Features and Details

[1]Impaired ability to adjust the body's position.
[2]Issues affecting the middle of the body.

Name	μ	σ	min-max
nms_d1	1.73	3.35	0-24
nms_d2	8.75	8.70	0-48
nms_d3	8.68	11.55	0-60
nms_d4	1.64	3.86	0-33
nms_d5	5.42	7.43	0-36
nms_d6	5.53	6.79	0-36
nms_d7	8.08	8.94	0-36
nms_d8	3.52	5.97	0-24
nms_d9	7.13	7.79	0-48
tremor	2.59	2.58	0-12
bradykin	2.40	1.41	0-6
rigidity	2.24	1.36	0-6
axial	3.25	2.68	0-12

Table 2: Descriptive Statistics

2 k-means

k-means clustering with $k = 4$ was tried. $k = 2, 3$ provided models that were too simplistic. $k = 5$ did not provide any new information, but rather just fragmented existing groups.

Table 3: Cluster statistics

Cluster	n
1	189
2	88
3	221
4	406

2.1 Decision tree

2.2 Interpretation of Clusters

2.2.1 Cluster summaries

Available in Figure 2. Error bar is standard error.

2.2.2 Interpretation

2.2.3 Statistical Significance Tests, $k = 4$

Using one-way ANOVA for multiple means, we reject the null hypothesis that the means are the same with $p < 0.05$ for every variable *except* `pdonset`.

Post-hoc analysis using Tukey's HSD:

```
age insignificant differences:
          diff       lwr       upr     p adj
3-1  0.9947808 -1.458184 3.4477455 0.7236845
4-1 -1.2838898 -3.464063 0.8962832 0.4284274
sex insignificant differences:
           diff        lwr        upr      p adj
2-1 -0.05044493 -0.2106093 0.10971941 0.84945412
4-1 -0.09897829 -0.2082638 0.01030726 0.09181043
3-2 -0.12633690 -0.2827741 0.03010026 0.16087872
4-2 -0.04853336 -0.1944676 0.09740091 0.82744866
4-3  0.07780354 -0.0259428 0.18154987 0.21607772
pdonset insignificant differences:
          diff        lwr      upr     p adj
2-1  2.9315777 -0.6232172 6.486373 0.1466742
3-1  1.7136632 -1.0153886 4.442715 0.3699280
4-1  0.7453932 -1.6801637 3.170950 0.8585776
3-2 -1.2179144 -4.6899860 2.254157 0.8033301
4-2 -2.1861845 -5.4251477 1.052779 0.3049434
4-3 -0.9682701 -3.2708860 1.334346 0.7004488
durat_pd insignificant differences:
          diff       lwr        upr        p adj
3-1 -0.7188824 -2.140915  0.7031499 5.624040e-01
cisitot insignificant differences:
          diff         lwr       upr        p adj
3-1  0.4942421  -0.4388731  1.427357 5.228228e-01
nms_d1 insignificant differences:
          diff        lwr        upr        p adj
4-3 -0.3798787 -0.9604053  0.2006478 3.325894e-01
nms_d4 insignificant differences:
          diff        lwr        upr        p adj
4-3 -0.3362459 -0.9972012  0.3247094 5.571480e-01
nms_d5 insignificant differences:
           diff        lwr         upr        p adj
4-3  -0.4117201  -1.743630   0.9201902 8.564409e-01
nms_d8 insignificant differences:
          diff        lwr        upr        p adj
4-3 -0.9953302 -2.1560641  0.1654036 1.220509e-01
nms_d9 insignificant differences:
          diff       lwr        upr      p adj
2-1  0.8708514 -1.297413 3.03911557 0.72966265
4-3 -1.3221920 -2.726684 0.08229967 0.07350641
tremor insignificant differences:
          diff         lwr        upr        p adj
4-1  0.3346105 -0.19270261  0.8619236 3.603863e-01
```

2.2.4 Ranked Features by Information Gain

Table 4: Features ranked by information gain

variable	information gain
bradykin	0.31574672
rigidity	0.29560018
nms_d2	0.24218407
cisitot	0.22920103
axial	0.22780750
nms_d3	0.20480570
nms_d9	0.15782743
nms_d7	0.15290569
nms_d5	0.14454931
nms_d6	0.14025139
nms_d1	0.13212756
tremor	0.10937168
nms_d4	0.10710526
nms_d8	0.10005480
durat_pd	0.02876190
age	0.02346158
sex	0.00000000
pdonset	0.00000000

2.2.5 Correlation Plots

Figure 3.

2.2.6 One vs all decision trees

Figures 4, 5, 6, 7.

3 Other Work

3.1 Bayesian Networks on Cluster 1

Figures 8, 9, and 10 show various learning algorithms.

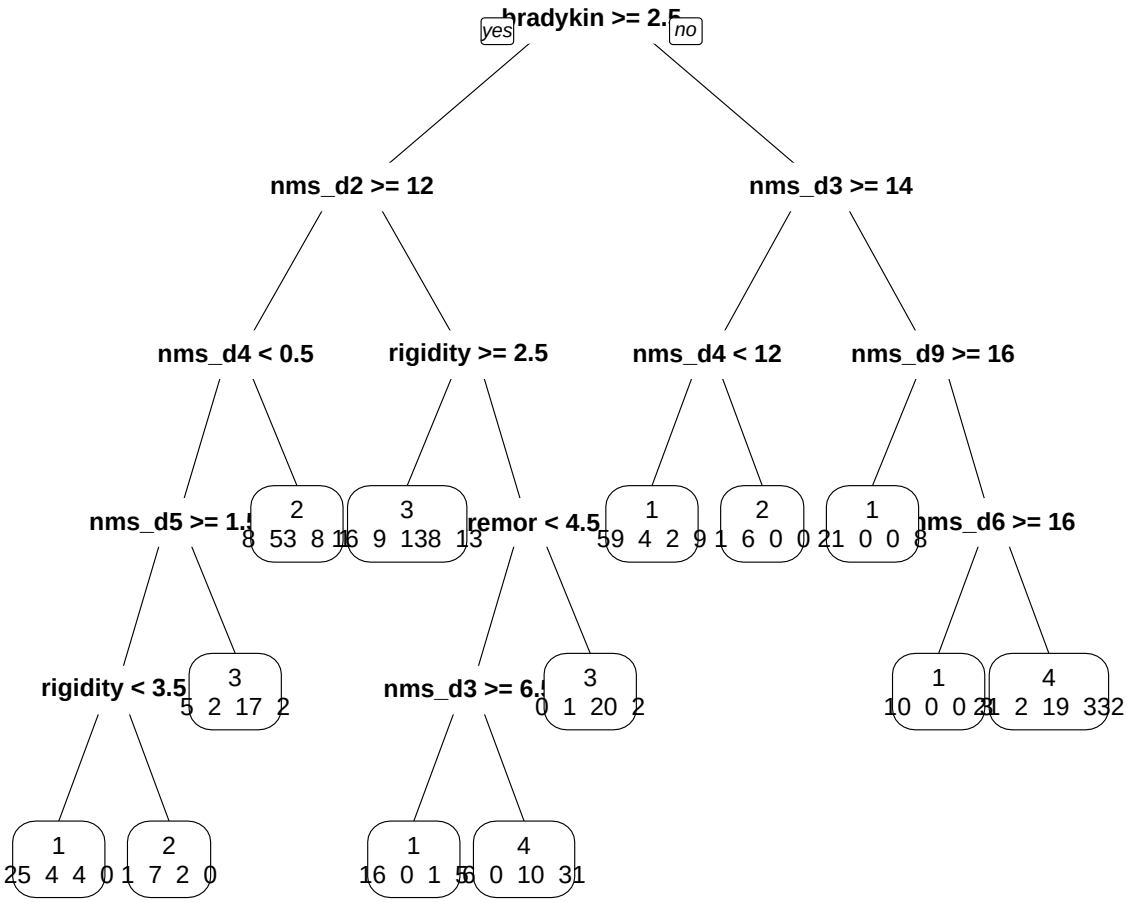

Figure 1: Decision Tree from k-means clustering, 4 clusters

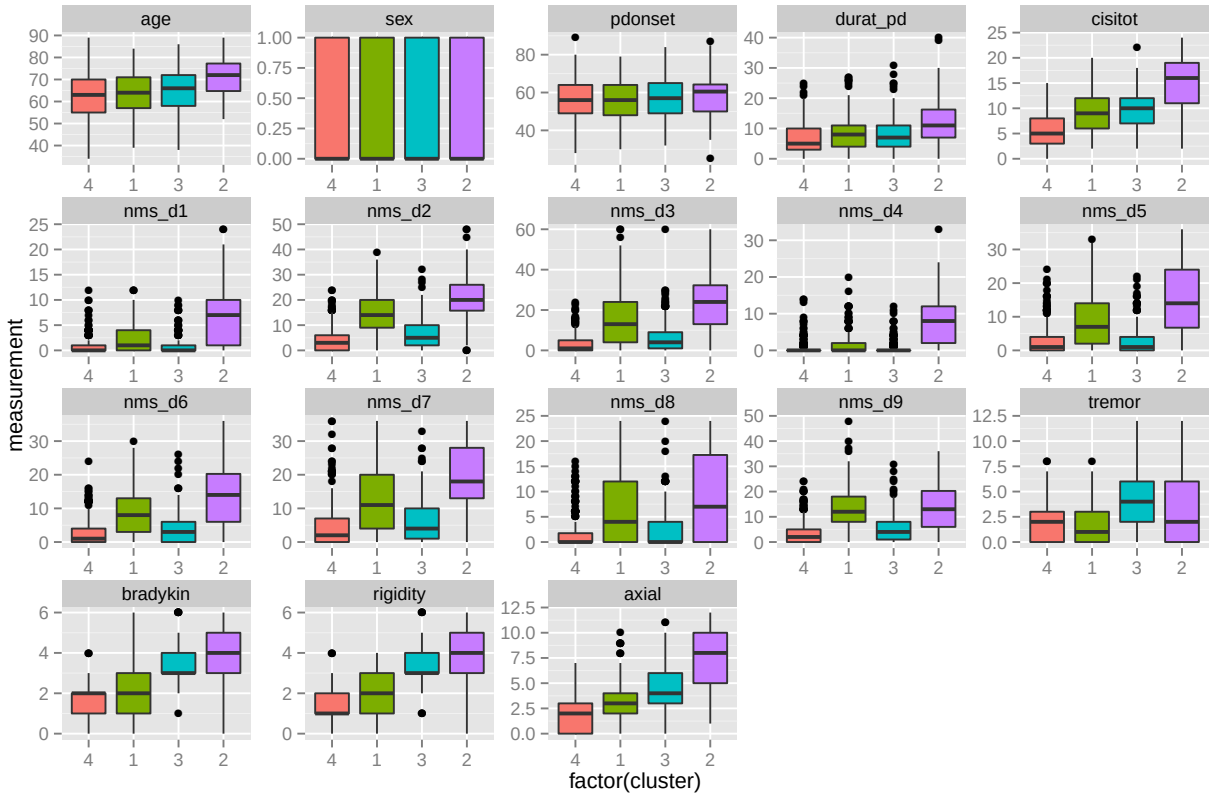

Figure 2: Cluster Summaries, $k = 4$

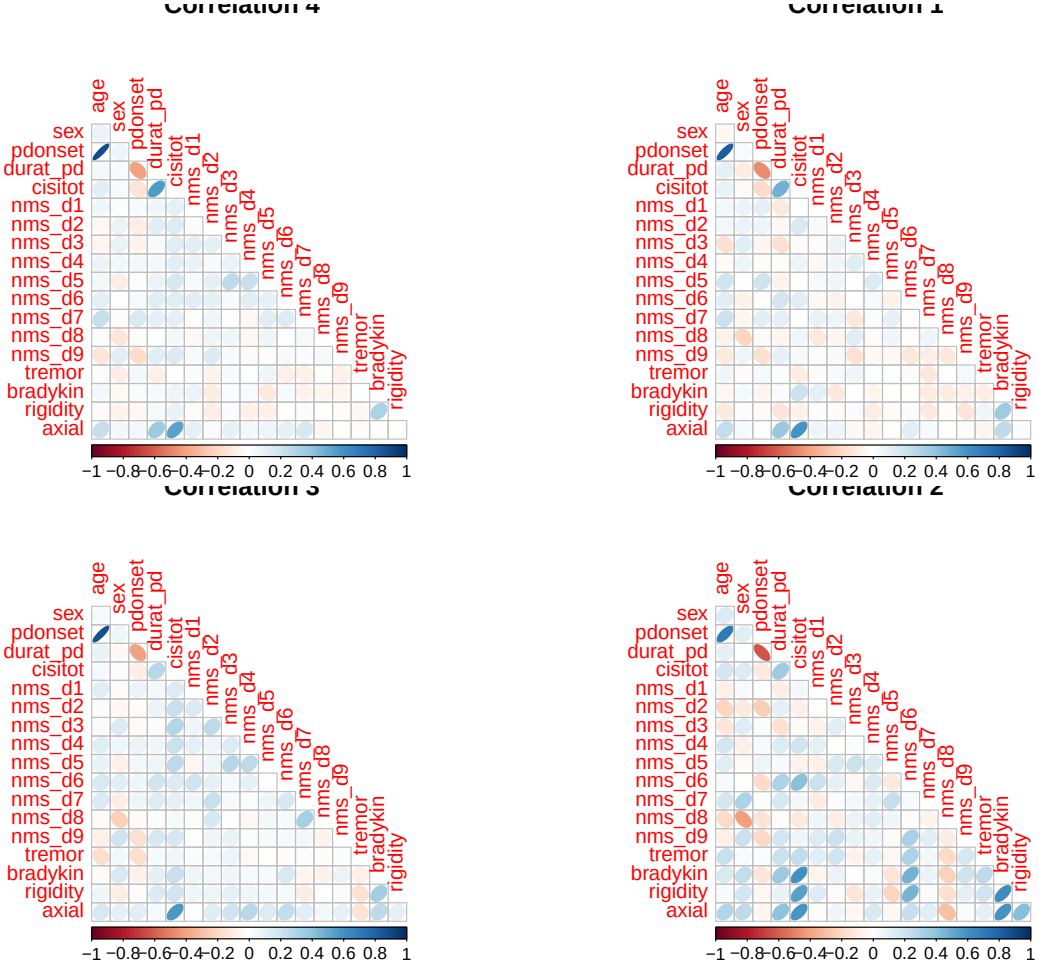

Figure 3: Correlation plots

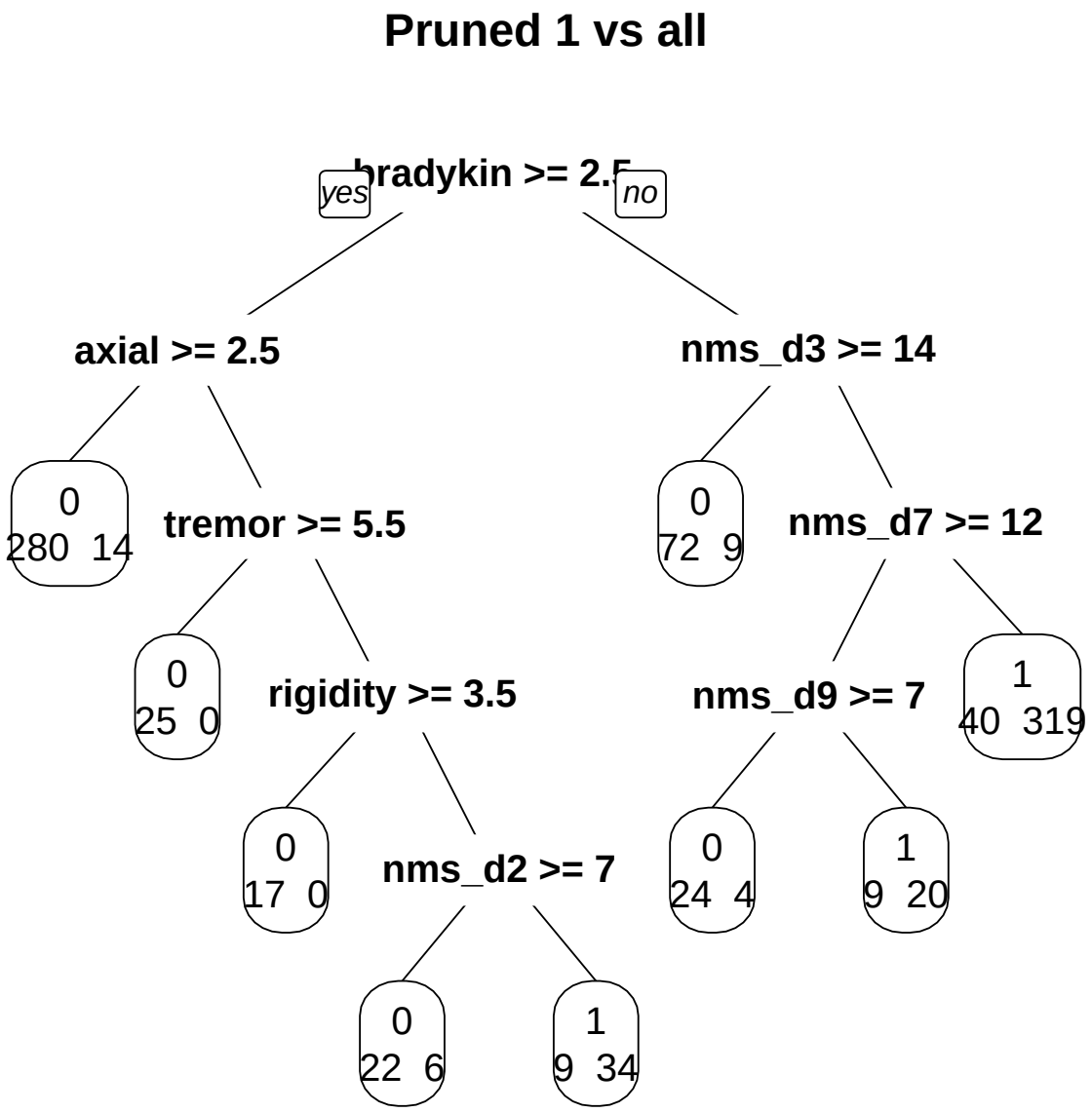

Figure 4: Cluster 1 vs all

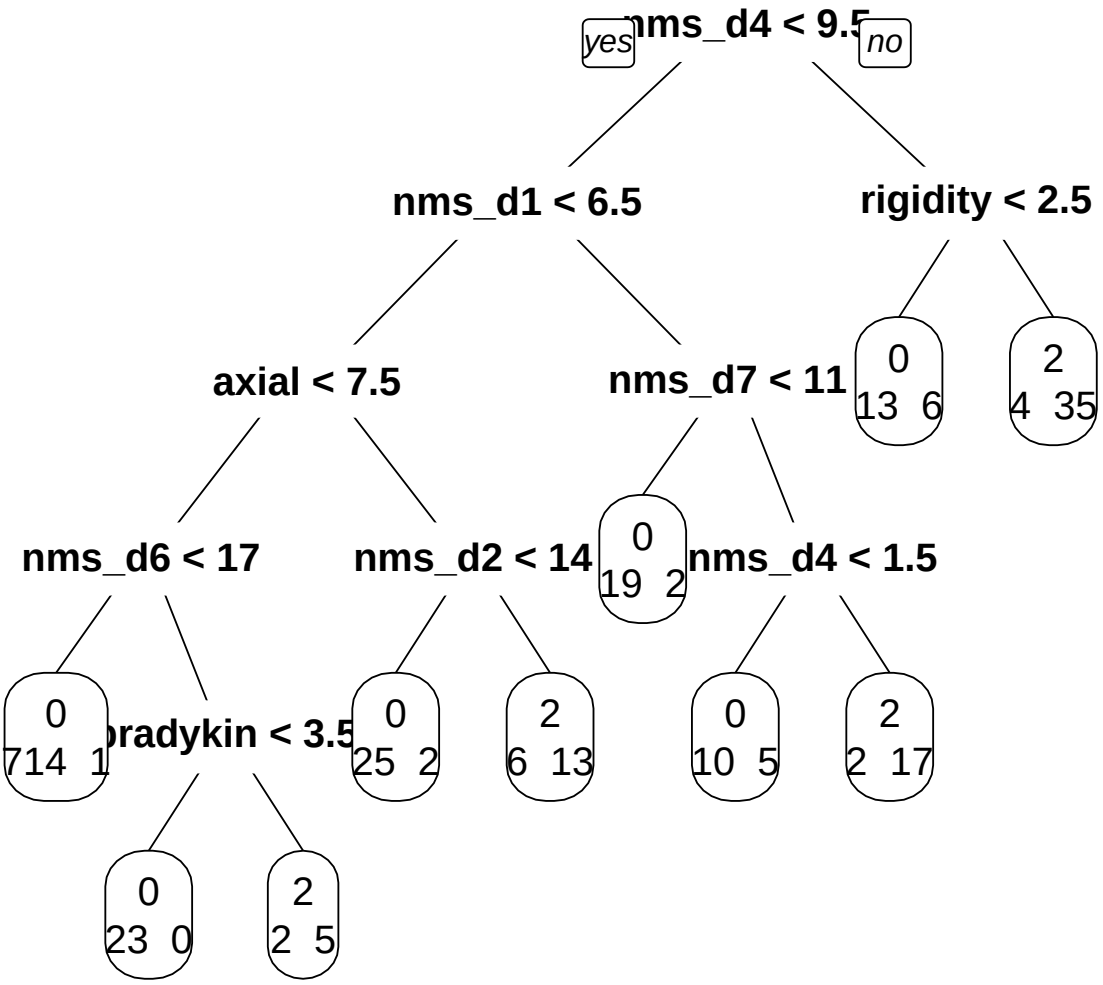

Figure 5: Cluster 2 vs all

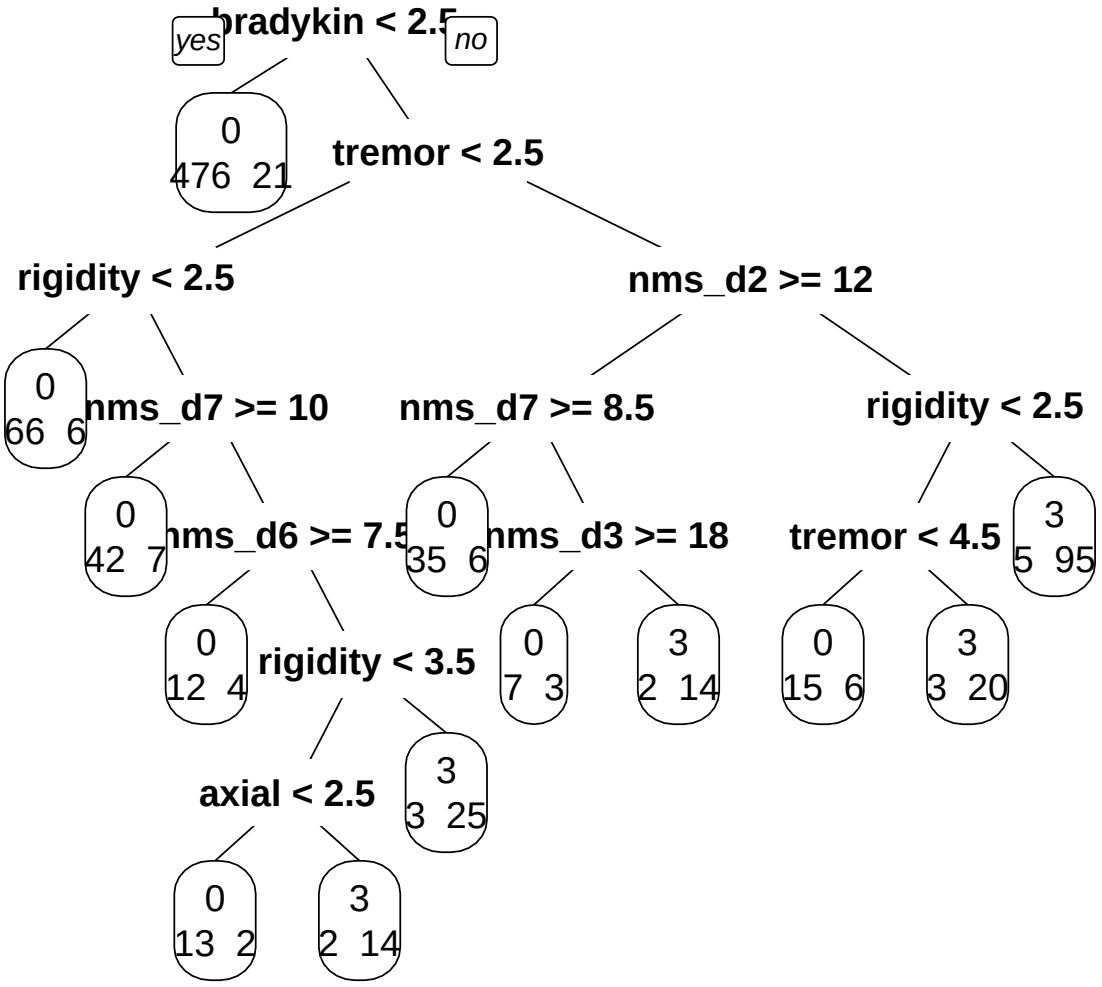

Figure 6: Cluster 3 vs all

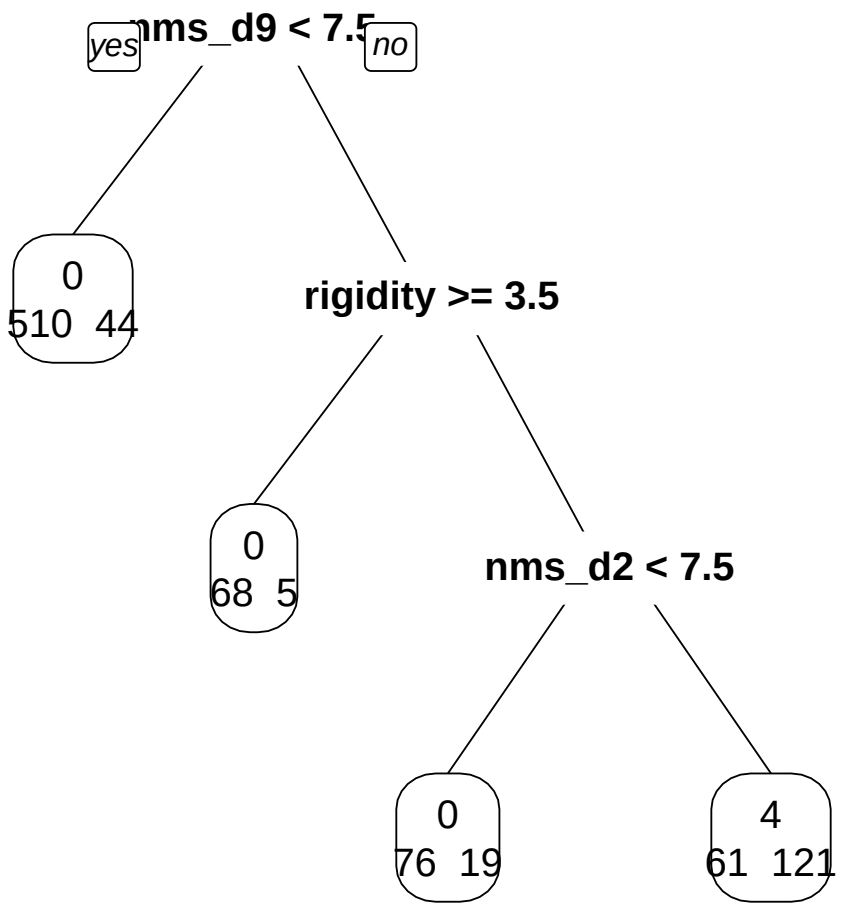

Figure 7: Cluster 4 vs all

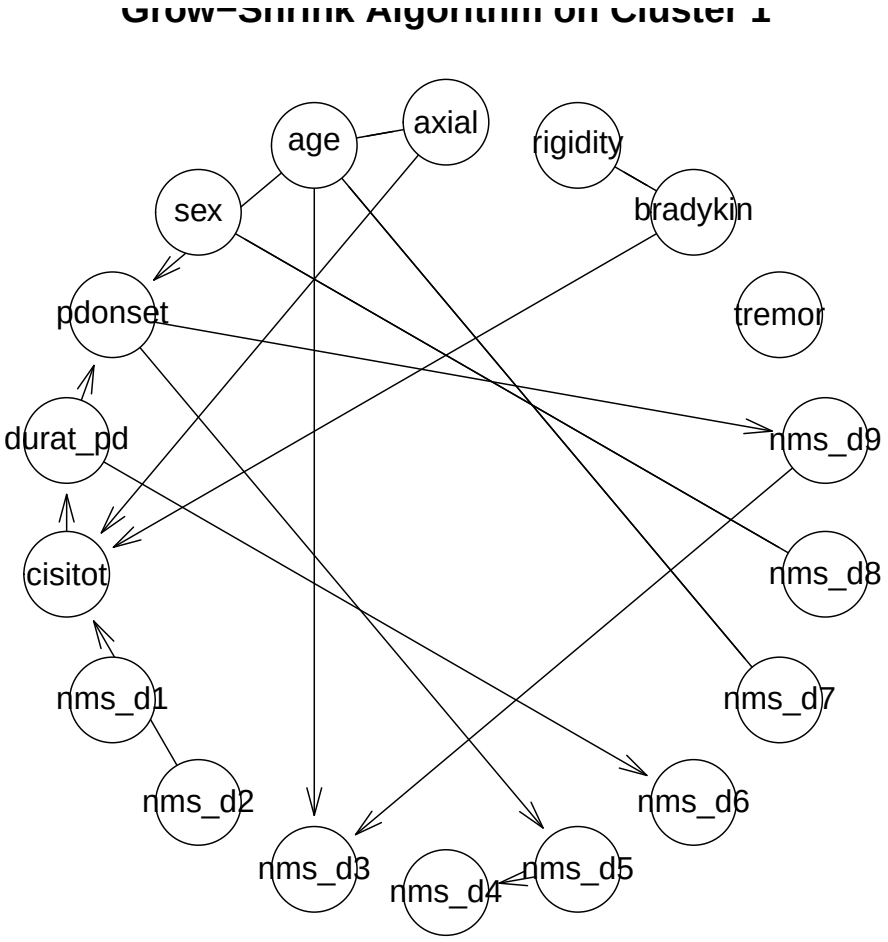

Figure 8: Grow Shrink Algorithm

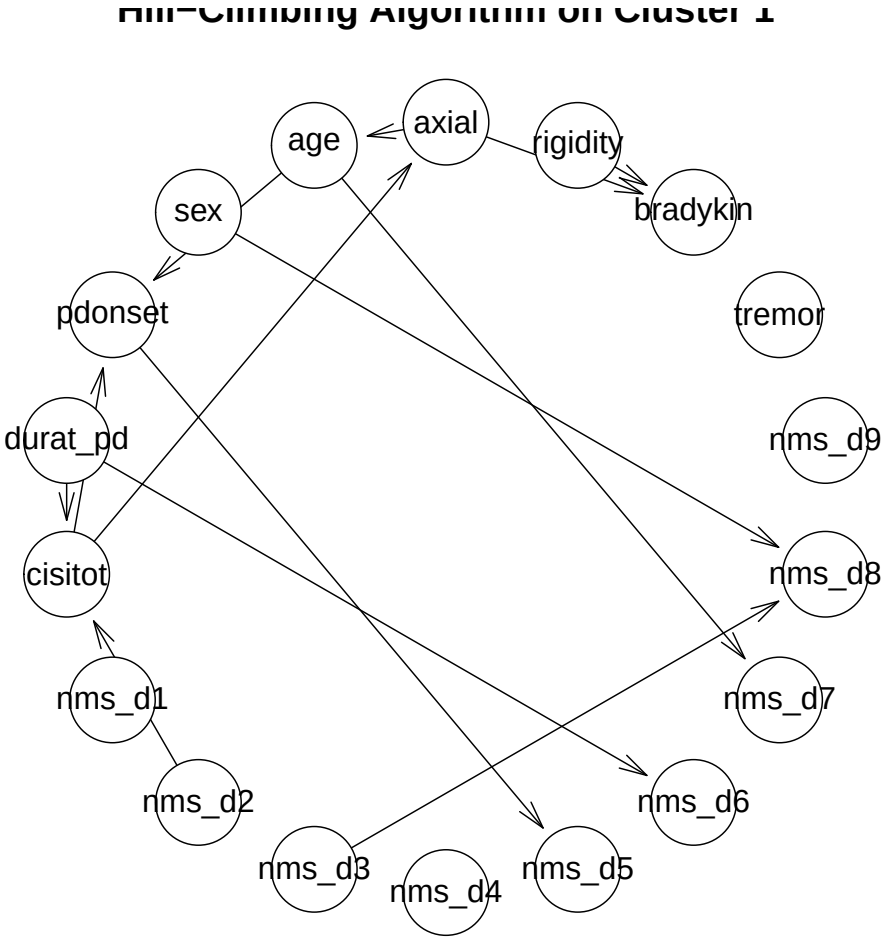

Figure 9: Hill Climbing Algorithm

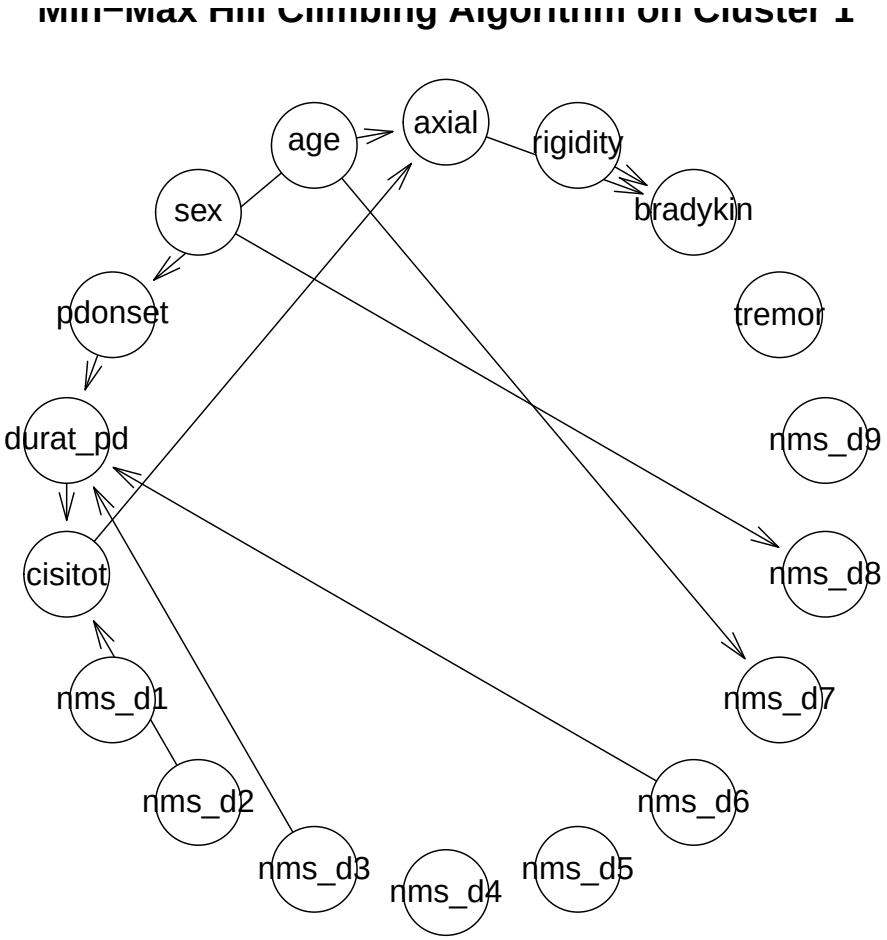

Figure 10: Min-Max Hill Climbing Algorithm