boruta_trials

May 28, 2020

[1]:
```
import numpy as np
import pandas as pd
from sklearn.ensemble import RandomForestClassifier
from boruta import BorutaPy
from sklearn import preprocessing
```

1 Transforming and Splitting Data

[2]:
```
df = pd.read_csv("data/combined_expression.csv")
df.head()
```

[2]:

	CELL_LINE_NAME	cluster	TSPAN6	TNMD	DPM1	SCYL3	C1orf112	\
0	1240123	2	8.319417	3.111183	9.643558	4.757258	3.919757	
1	1240131	1	7.611268	2.704739	10.276079	3.650299	3.481567	
2	1240132	1	7.678658	2.845781	10.180954	3.573048	3.431235	
3	1240134	1	3.265063	3.063746	10.490285	3.340791	3.676912	
4	1240140	1	7.090138	2.988043	10.264692	4.119555	3.432585	

	FGR	CFH	FUCA2	…	C6orf10	TMEM225	NOTCH4	PBX2	\
0	3.602185	3.329644	9.076950	…	3.085394	3.462811	3.339030	4.614897	
1	3.145538	3.565127	7.861068	…	2.801456	2.985889	3.180068	5.415729	
2	3.090781	4.116643	8.121190	…	2.934962	2.952937	3.164655	5.707506	
3	3.512821	3.873922	8.790851	…	3.041839	3.398847	3.106710	5.773963	
4	3.308033	3.318371	6.927761	…	3.028787	3.225982	3.275820	5.334283	

	AGER	RNF5	AGPAT1	DFNB59	PRRT1	FKBPL
0	3.395845	3.419193	3.971646	3.729310	3.320022	6.447316
1	3.299858	3.028414	3.877889	3.911516	3.379405	4.729557
2	3.434295	2.961345	4.272194	3.085696	3.002557	5.653588
3	3.412641	3.136110	4.422262	3.522122	3.509437	5.953242
4	3.864678	3.259242	3.840581	5.809553	3.674587	5.577503

[5 rows x 16384 columns]

[3]:
```
features = [f for f in  df.columns if f not in ['CELL_LINE_NAME', 'cluster']]
len(features)
```

[3]: 16382

[4]:
```
X = df[features].values
Y = df['cluster'].values.ravel()
```

[5]:
```
min_max_scaler = preprocessing.MinMaxScaler()
X = min_max_scaler.fit_transform(X)
```

[7]:
```
# max_depth of tree advised on Boruta Github to be ~3-7
rf = RandomForestClassifier(n_jobs=-1, class_weight='balanced', max_depth=5)
boruta_feature_selector = BorutaPy(rf, n_estimators='auto', verbose=2,␣
 ↪random_state=1, max_iter=100)
boruta_feature_selector.fit(X, Y)
```

```
Iteration:      1 / 100
Confirmed:      0
Tentative:      16382
Rejected:       0
Iteration:      2 / 100
Confirmed:      0
Tentative:      16382
Rejected:       0
Iteration:      3 / 100
Confirmed:      0
Tentative:      16382
Rejected:       0
Iteration:      4 / 100
Confirmed:      0
Tentative:      16382
Rejected:       0
Iteration:      5 / 100
Confirmed:      0
Tentative:      16382
Rejected:       0
Iteration:      6 / 100
Confirmed:      0
Tentative:      16382
Rejected:       0
Iteration:      7 / 100
Confirmed:      0
Tentative:      16382
Rejected:       0
Iteration:      8 / 100
Confirmed:      0
Tentative:      1479
Rejected:       14903
Iteration:      9 / 100
```

```
Confirmed:      255
Tentative:      1224
Rejected:       14903
Iteration:      10 / 100
Confirmed:      255
Tentative:      1224
Rejected:       14903
Iteration:      11 / 100
Confirmed:      255
Tentative:      1224
Rejected:       14903
Iteration:      12 / 100
Confirmed:      265
Tentative:      848
Rejected:       15269
Iteration:      13 / 100
Confirmed:      265
Tentative:      848
Rejected:       15269
Iteration:      14 / 100
Confirmed:      265
Tentative:      848
Rejected:       15269
Iteration:      15 / 100
Confirmed:      265
Tentative:      848
Rejected:       15269
Iteration:      16 / 100
Confirmed:      267
Tentative:      706
Rejected:       15409
Iteration:      17 / 100
Confirmed:      267
Tentative:      706
Rejected:       15409
Iteration:      18 / 100
Confirmed:      267
Tentative:      706
Rejected:       15409
Iteration:      19 / 100
Confirmed:      269
Tentative:      600
Rejected:       15513
Iteration:      20 / 100
Confirmed:      269
Tentative:      600
Rejected:       15513
Iteration:      21 / 100
```

```
Confirmed:      269
Tentative:      600
Rejected:       15513
Iteration:      22 / 100
Confirmed:      270
Tentative:      506
Rejected:       15606
Iteration:      23 / 100
Confirmed:      270
Tentative:      506
Rejected:       15606
Iteration:      24 / 100
Confirmed:      270
Tentative:      506
Rejected:       15606
Iteration:      25 / 100
Confirmed:      270
Tentative:      506
Rejected:       15606
Iteration:      26 / 100
Confirmed:      270
Tentative:      445
Rejected:       15667
Iteration:      27 / 100
Confirmed:      270
Tentative:      445
Rejected:       15667
Iteration:      28 / 100
Confirmed:      270
Tentative:      445
Rejected:       15667
Iteration:      29 / 100
Confirmed:      272
Tentative:      383
Rejected:       15727
Iteration:      30 / 100
Confirmed:      272
Tentative:      383
Rejected:       15727
Iteration:      31 / 100
Confirmed:      272
Tentative:      383
Rejected:       15727
Iteration:      32 / 100
Confirmed:      274
Tentative:      329
Rejected:       15779
Iteration:      33 / 100
```

```
Confirmed:      274
Tentative:      329
Rejected:       15779
Iteration:      34 / 100
Confirmed:      276
Tentative:      285
Rejected:       15821
Iteration:      35 / 100
Confirmed:      276
Tentative:      285
Rejected:       15821
Iteration:      36 / 100
Confirmed:      276
Tentative:      285
Rejected:       15821
Iteration:      37 / 100
Confirmed:      278
Tentative:      242
Rejected:       15862
Iteration:      38 / 100
Confirmed:      278
Tentative:      242
Rejected:       15862
Iteration:      39 / 100
Confirmed:      278
Tentative:      242
Rejected:       15862
Iteration:      40 / 100
Confirmed:      282
Tentative:      220
Rejected:       15880
Iteration:      41 / 100
Confirmed:      282
Tentative:      220
Rejected:       15880
Iteration:      42 / 100
Confirmed:      282
Tentative:      220
Rejected:       15880
Iteration:      43 / 100
Confirmed:      284
Tentative:      193
Rejected:       15905
Iteration:      44 / 100
Confirmed:      284
Tentative:      193
Rejected:       15905
Iteration:      45 / 100
```

```
Confirmed:      284
Tentative:      193
Rejected:       15905
Iteration:      46 / 100
Confirmed:      284
Tentative:      184
Rejected:       15914
Iteration:      47 / 100
Confirmed:      284
Tentative:      184
Rejected:       15914
Iteration:      48 / 100
Confirmed:      284
Tentative:      184
Rejected:       15914
Iteration:      49 / 100
Confirmed:      286
Tentative:      182
Rejected:       15914
Iteration:      50 / 100
Confirmed:      286
Tentative:      179
Rejected:       15917
Iteration:      51 / 100
Confirmed:      290
Tentative:      169
Rejected:       15923
Iteration:      52 / 100
Confirmed:      290
Tentative:      169
Rejected:       15923
Iteration:      53 / 100
Confirmed:      290
Tentative:      169
Rejected:       15923
Iteration:      54 / 100
Confirmed:      291
Tentative:      153
Rejected:       15938
Iteration:      55 / 100
Confirmed:      291
Tentative:      153
Rejected:       15938
Iteration:      56 / 100
Confirmed:      291
Tentative:      153
Rejected:       15938
Iteration:      57 / 100
```

```
Confirmed:      292
Tentative:      150
Rejected:       15940
Iteration:      58 / 100
Confirmed:      292
Tentative:      150
Rejected:       15940
Iteration:      59 / 100
Confirmed:      295
Tentative:      147
Rejected:       15940
Iteration:      60 / 100
Confirmed:      295
Tentative:      147
Rejected:       15940
Iteration:      61 / 100
Confirmed:      295
Tentative:      147
Rejected:       15940
Iteration:      62 / 100
Confirmed:      296
Tentative:      141
Rejected:       15945
Iteration:      63 / 100
Confirmed:      296
Tentative:      141
Rejected:       15945
Iteration:      64 / 100
Confirmed:      296
Tentative:      141
Rejected:       15945
Iteration:      65 / 100
Confirmed:      296
Tentative:      139
Rejected:       15947
Iteration:      66 / 100
Confirmed:      296
Tentative:      139
Rejected:       15947
Iteration:      67 / 100
Confirmed:      296
Tentative:      133
Rejected:       15953
Iteration:      68 / 100
Confirmed:      296
Tentative:      133
Rejected:       15953
Iteration:      69 / 100
```

```
Confirmed:      296
Tentative:      133
Rejected:       15953
Iteration:      70 / 100
Confirmed:      296
Tentative:      128
Rejected:       15958
Iteration:      71 / 100
Confirmed:      296
Tentative:      128
Rejected:       15958
Iteration:      72 / 100
Confirmed:      296
Tentative:      122
Rejected:       15964
Iteration:      73 / 100
Confirmed:      296
Tentative:      122
Rejected:       15964
Iteration:      74 / 100
Confirmed:      296
Tentative:      122
Rejected:       15964
Iteration:      75 / 100
Confirmed:      296
Tentative:      114
Rejected:       15972
Iteration:      76 / 100
Confirmed:      296
Tentative:      114
Rejected:       15972
Iteration:      77 / 100
Confirmed:      296
Tentative:      111
Rejected:       15975
Iteration:      78 / 100
Confirmed:      296
Tentative:      111
Rejected:       15975
Iteration:      79 / 100
Confirmed:      296
Tentative:      111
Rejected:       15975
Iteration:      80 / 100
Confirmed:      297
Tentative:      109
Rejected:       15976
Iteration:      81 / 100
```

```
Confirmed:      297
Tentative:      109
Rejected:       15976
Iteration:      82 / 100
Confirmed:      297
Tentative:      109
Rejected:       15976
Iteration:      83 / 100
Confirmed:      299
Tentative:      107
Rejected:       15976
Iteration:      84 / 100
Confirmed:      299
Tentative:      107
Rejected:       15976
Iteration:      85 / 100
Confirmed:      299
Tentative:      103
Rejected:       15980
Iteration:      86 / 100
Confirmed:      299
Tentative:      103
Rejected:       15980
Iteration:      87 / 100
Confirmed:      299
Tentative:      103
Rejected:       15980
Iteration:      88 / 100
Confirmed:      299
Tentative:      100
Rejected:       15983
Iteration:      89 / 100
Confirmed:      299
Tentative:      100
Rejected:       15983
Iteration:      90 / 100
Confirmed:      300
Tentative:      97
Rejected:       15985
Iteration:      91 / 100
Confirmed:      300
Tentative:      97
Rejected:       15985
Iteration:      92 / 100
Confirmed:      300
Tentative:      97
Rejected:       15985
Iteration:      93 / 100
```

```
Confirmed:      300
Tentative:      96
Rejected:       15986
Iteration:      94 / 100
Confirmed:      300
Tentative:      96
Rejected:       15986
Iteration:      95 / 100
Confirmed:      300
Tentative:      95
Rejected:       15987
Iteration:      96 / 100
Confirmed:      300
Tentative:      95
Rejected:       15987
Iteration:      97 / 100
Confirmed:      300
Tentative:      95
Rejected:       15987
Iteration:      98 / 100
Confirmed:      301
Tentative:      93
Rejected:       15988
Iteration:      99 / 100
Confirmed:      301
Tentative:      93
Rejected:       15988


BorutaPy finished running.

Iteration:      100 / 100
Confirmed:      301
Tentative:      24
Rejected:       15988
```

[7]:
```
BorutaPy(estimator=RandomForestClassifier(class_weight='balanced', max_depth=5,
                                          n_estimators=561, n_jobs=-1,
                                          random_state=RandomState(MT19937) at
0x1A1E8C5990),
         n_estimators='auto', random_state=RandomState(MT19937) at 0x1A1E8C5990,
         verbose=2)
```

[8]:
```
# check selected features - first 5 features are selected
boruta_feature_selector.support_
```

[8]:
```
array([False, False, False, …, False, False, False])
```

[9]:
```
# check ranking of features
boruta_feature_selector.ranking_
```

[9]:
```
array([  244,  4614,  4220, …, 10181, 11398, 15509])
```

[10]:
```
X_filtered = boruta_feature_selector.transform(X)
X_filtered.shape
```

[10]:
```
(541, 301)
```

[11]:
```
final_features = list()
indices = np.where(boruta_feature_selector.support_ == True)
for x in np.nditer(indices):
    final_features.append(features[x])
final_features
```

[11]:
```
['FAM214B',
 'ITGA3',
 'TNFRSF12A',
 'ALDH3B1',
 'RHBDF1',
 'CYTH3',
 'HFE',
 'MVP',
 'GPRC5A',
 'CCDC88C',
 'WWTR1',
 'SAMD4A',
 'VIM',
 'CTNNA1',
 'POLR2B',
 'DTNBP1',
 'VAMP3',
 'BCAR1',
 'FOXC1',
 'DCBLD2',
 'NCKAP1',
 'GPC1',
 'CTSA',
 'SUGP2',
 'SNX24',
 'PTPN21',
 'DAZAP1',
 'ACTN1',
 'PPP2R3A',
 'IGF2BP2',
 'NTN4',
```

'NUAK1',
'SEMA3C',
'RASAL2',
'FNDC3B',
'FOSL2',
'PLD1',
'RBMS2',
'EDN1',
'ITGB5',
'SMAP2',
'CD59',
'CTTN',
'EPB41L1',
'SNX5',
'KDM2B',
'PXN',
'LAMB1',
'TBC1D2',
'CDC7',
'KDELR3',
'VRK1',
'NOP56',
'POLA1',
'PLS3',
'CORO1A',
'GABPB1',
'TJP1',
'UBR5',
'CLASRP',
'RASAL3',
'SUGP1',
'TFPI2',
'OGDH',
'CAV2',
'CAV1',
'MET',
'HIBADH',
'SERPINE1',
'EZH2',
'PLEKHA1',
'DKK1',
'BLMH',
'ABCC3',
'DUSP3',
'TNFAIP1',
'SH3D19',
'CCND1',

'PRPF19',
'ARHGEF17',
'CPSF6',
'RPS12',
'AMOTL2',
'FHL2',
'RND3',
'EPAS1',
'RPL22',
'ERRFI1',
'F3',
'ARID1A',
'RAB32',
'MYB',
'CTGF',
'LTBP2',
'AVPI1',
'RCL1',
'TGFBI',
'B4GALT4',
'KHDRBS1',
'CXCR4',
'KIAA1191',
'CALD1',
'GIPC1',
'ARHGAP9',
'NCKAP1L',
'PREX1',
'SDC4',
'AHNAK',
'UPF3B',
'X06.Sep',
'SLC25A19',
'RRAS',
'TRAP1',
'L3HYPDH',
'HSPA2',
'YEATS4',
'PPAT',
'KRI1',
'ACTN4',
'LAMA5',
'CLIP1',
'SLC35D2',
'LGALS3',
'NUP210',
'PNISR',

'RIN2',
'VHL',
'EMP1',
'X.14',
'RNF138',
'ANXA1',
'NT5E',
'RAB11FIP5',
'LAMC1',
'TMBIM1',
'TNS3',
'SRSF1',
'FAM129B',
'RPS6',
'IER3',
'MDC1',
'YAP1',
'THBS1',
'UACA',
'BCAR3',
'ARHGAP29',
'MYOF',
'ITGAV',
'DIRC2',
'PHLDA1',
'SDSL',
'TCHP',
'HNRNPA1L2',
'SLAIN1',
'TGFB1I1',
'KIFC3',
'NARF',
'FKBP10',
'IFITM3',
'RPS11',
'EPHA2',
'HSPG2',
'CYR61',
'RGS16',
'LBR',
'GULP1',
'CTDSPL',
'NCEH1',
'LPP',
'KLHL8',
'TIFA',
'RPS3A',

```
'OSMR',
'PLK2',
'FBXL17',
'EGFR',
'ZNF92',
'NONO',
'OGT',
'ZNF711',
'RBMX',
'GSN',
'SLC43A1',
'SERPINH1',
'ITGB1',
'LATS2',
'PRSS23',
'DOCK1',
'CRIM1',
'QDPR',
'ASAP2',
'WWC2',
'BAG3',
'CAST',
'RBMS1',
'ZDHHC7',
'RHOC',
'DCK',
'SH3RF2',
'RBPMS',
'MYO1E',
'EPB41',
'PAXBP1',
'SAFB',
'LMNA',
'SQSTM1',
'SNX7',
'CAPN2',
'S100A11',
'CLDN1',
'FSTL1',
'CLASP2',
'CXCL1',
'GNL3',
'SGMS2',
'ZNF589',
'CAMKV',
'HMGB2',
'ITGA2',
```

```
'GPX8',
'RAD21',
'MICALL2',
'STRBP',
'UBTD1',
'HTRA1',
'PRKCB',
'CCDC68',
'NNMT',
'SMAD3',
'AXL',
'C19orf33',
'EEF2',
'KRT80',
'IGFBP6',
'VASN',
'RHOH',
'RAB31',
'ADAM9',
'PPIC',
'PXDC1',
'ATF5',
'UGP2',
'TM4SF1',
'PUSL1',
'BCL2',
'THOP1',
'LAMB2',
'GNG12',
'PHF8',
'DAG1',
'MSL2',
'RPL15',
'PDIK1L',
'FOSL1',
'ATAD5',
'CLK2',
'CCDC101',
'PTRF',
'CD151',
'ALS2CL',
'PLEC',
'BEND3',
'NQO1',
'PHLDA2',
'ZNF708',
'EXT1',
```

```
 'ANXA2',
 'RBM10',
 'UPP1',
 'MED12',
 'SV2B',
 'AHNAK2',
 'BCL9L',
 'UBE2H',
 'MIR22HG',
 'MAPT',
 'KCNJ11',
 'USP7',
 'IER5L',
 'C22orf34',
 'C15orf52',
 'S100A16',
 'RPL14',
 'BEND4',
 'S100A13',
 'SPATS2L',
 'C20orf96',
 'NACA',
 'HDAC2',
 'ANXA4',
 'LRRC8B',
 'BLM',
 'PARVA',
 'FAM114A1',
 'S100A10',
 'ATAD3A',
 'MYO1C',
 'S100A6',
 'ASPH',
 'PAPSS2',
 'SHISA4']
```

[12]:
```
s_feats = pd.DataFrame(final_features)
s_feats.to_csv('cleaned/boruta.csv', index=False)
```

[ ]:
```

```