boruta_trials

May 26, 2020

[91]:
```
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

[92]:
```
df = pd.read_csv("data/combined_expression.csv")
df.head()
```

[92]:
```
   CELL_LINE_NAME  classification    TSPAN6      TNMD       DPM1     SCYL3  \
0         1240121               1  6.419526  3.182094   9.320548  3.759654
1         1240122               2  7.646494  2.626819  10.153853  3.564755
2         1240123               1  8.319417  3.111183   9.643558  4.757258
3         1240124               1  9.006994  3.028173   9.686700  4.280504
4         1240127               1  7.985676  2.694729  10.676134  4.159685

   C1orf112       FGR       CFH     FUCA2  …   COL15A1   C6orf10   TMEM225  \
0  3.802619  3.215753  4.698729  7.873672  …  3.245454  2.953508  3.543429
1  3.942749  3.290760  3.551675  8.252413  …  2.786709  3.077382  3.728232
2  3.919757  3.602185  3.329644  9.076950  …  3.459089  3.085394  3.462811
3  3.147646  3.188881  3.293807  8.678790  …  2.835403  2.960303  3.415083
4  3.804637  3.481942  3.111261  7.555407  …  2.896523  2.849899  3.480114

     NOTCH4      PBX2      AGER      RNF5    AGPAT1    DFNB59     PRRT1
0  3.352022  4.672310  3.641128  3.135310  3.737072  3.450927  3.168800
1  3.208882  4.586840  3.395654  3.586800  3.519128  3.115323  3.051645
2  3.339030  4.614897  3.395845  3.419193  3.971646  3.729310  3.320022
3  3.290171  4.770123  3.400821  3.383734  3.798107  2.822404  3.297547
4  3.226128  5.832710  3.612179  3.347095  4.457963  5.198524  4.553586

[5 rows x 16383 columns]
```

[99]:
```
features = [f for f in  df.columns if f not in ['CELL_LINE_NAME',␣
 ↪'classification']]
```

```
len(features)
```

```
[99]: 16381
```

```
[100]: X = df[features].values
       Y = df['classification'].values.ravel()
```

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

```
[102]: # max_depth of tree advised on Boruta Github to be ~3-7
       rf = RandomForestClassifier(n_jobs=-1, class_weight='balanced', max_depth=3)
       boruta_feature_selector = BorutaPy(rf, n_estimators='auto', verbose=2,␣
        ↪random_state=1, perc=99, max_iter=50)
       boruta_feature_selector.fit(X, Y)
```

```
Iteration:      1 / 50
Confirmed:      0
Tentative:      16381
Rejected:       0
Iteration:      2 / 50
Confirmed:      0
Tentative:      16381
Rejected:       0
Iteration:      3 / 50
Confirmed:      0
Tentative:      16381
Rejected:       0
Iteration:      4 / 50
Confirmed:      0
Tentative:      16381
Rejected:       0
Iteration:      5 / 50
Confirmed:      0
Tentative:      16381
Rejected:       0
Iteration:      6 / 50
Confirmed:      0
Tentative:      16381
Rejected:       0
Iteration:      7 / 50
Confirmed:      0
Tentative:      16381
Rejected:       0
Iteration:      8 / 50
Confirmed:      0
Tentative:      1424
```

```
Rejected:       14957
Iteration:      9 / 50
Confirmed:      207
Tentative:      1217
Rejected:       14957
Iteration:      10 / 50
Confirmed:      207
Tentative:      1217
Rejected:       14957
Iteration:      11 / 50
Confirmed:      207
Tentative:      1217
Rejected:       14957
Iteration:      12 / 50
Confirmed:      239
Tentative:      797
Rejected:       15345
Iteration:      13 / 50
Confirmed:      239
Tentative:      797
Rejected:       15345
Iteration:      14 / 50
Confirmed:      239
Tentative:      797
Rejected:       15345
Iteration:      15 / 50
Confirmed:      239
Tentative:      797
Rejected:       15345
Iteration:      16 / 50
Confirmed:      256
Tentative:      654
Rejected:       15471
Iteration:      17 / 50
Confirmed:      256
Tentative:      654
Rejected:       15471
Iteration:      18 / 50
Confirmed:      256
Tentative:      654
Rejected:       15471
Iteration:      19 / 50
Confirmed:      263
Tentative:      546
Rejected:       15572
Iteration:      20 / 50
Confirmed:      263
Tentative:      546
```

```
Rejected:       15572
Iteration:      21 / 50
Confirmed:      263
Tentative:      546
Rejected:       15572
Iteration:      22 / 50
Confirmed:      268
Tentative:      473
Rejected:       15640
Iteration:      23 / 50
Confirmed:      268
Tentative:      473
Rejected:       15640
Iteration:      24 / 50
Confirmed:      268
Tentative:      473
Rejected:       15640
Iteration:      25 / 50
Confirmed:      268
Tentative:      473
Rejected:       15640
Iteration:      26 / 50
Confirmed:      273
Tentative:      435
Rejected:       15673
Iteration:      27 / 50
Confirmed:      273
Tentative:      435
Rejected:       15673
Iteration:      28 / 50
Confirmed:      273
Tentative:      435
Rejected:       15673
Iteration:      29 / 50
Confirmed:      275
Tentative:      401
Rejected:       15705
Iteration:      30 / 50
Confirmed:      275
Tentative:      401
Rejected:       15705
Iteration:      31 / 50
Confirmed:      275
Tentative:      401
Rejected:       15705
Iteration:      32 / 50
Confirmed:      276
Tentative:      376
```

```
Rejected:       15729
Iteration:      33 / 50
Confirmed:      276
Tentative:      376
Rejected:       15729
Iteration:      34 / 50
Confirmed:      277
Tentative:      348
Rejected:       15756
Iteration:      35 / 50
Confirmed:      277
Tentative:      348
Rejected:       15756
Iteration:      36 / 50
Confirmed:      277
Tentative:      348
Rejected:       15756
Iteration:      37 / 50
Confirmed:      278
Tentative:      329
Rejected:       15774
Iteration:      38 / 50
Confirmed:      278
Tentative:      329
Rejected:       15774
Iteration:      39 / 50
Confirmed:      278
Tentative:      329
Rejected:       15774
Iteration:      40 / 50
Confirmed:      279
Tentative:      328
Rejected:       15774
Iteration:      41 / 50
Confirmed:      279
Tentative:      319
Rejected:       15783
Iteration:      42 / 50
Confirmed:      279
Tentative:      319
Rejected:       15783
Iteration:      43 / 50
Confirmed:      280
Tentative:      311
Rejected:       15790
Iteration:      44 / 50
Confirmed:      280
Tentative:      311
```

```
Rejected:       15790
Iteration:      45 / 50
Confirmed:      280
Tentative:      311
Rejected:       15790
Iteration:      46 / 50
Confirmed:      280
Tentative:      300
Rejected:       15801
Iteration:      47 / 50
Confirmed:      280
Tentative:      300
Rejected:       15801
Iteration:      48 / 50
Confirmed:      280
Tentative:      300
Rejected:       15801
Iteration:      49 / 50
Confirmed:      280
Tentative:      291
Rejected:       15810


BorutaPy finished running.

Iteration:      50 / 50
Confirmed:      280
Tentative:      33
Rejected:       15810
```

```
[102]: BorutaPy(alpha=0.05,
                estimator=RandomForestClassifier(bootstrap=True, ccp_alpha=0.0,
                                                 class_weight='balanced',
                                                 criterion='gini', max_depth=3,
                                                 max_features='auto',
                                                 max_leaf_nodes=None, max_samples=None,
                                                 min_impurity_decrease=0.0,
                                                 min_impurity_split=None,
                                                 min_samples_leaf=1,
                                                 min_samples_split=2,
                                                 min_weight_fraction_leaf=0.0,
                                                 n_estimators=1135, n_jobs=-1,
                                                 oob_score=False,
                                                 random_state=RandomState(MT19937) at
       0x1A98C33678,
                                                 verbose=0, warm_start=False),
                max_iter=50, n_estimators='auto', perc=99,
                random_state=RandomState(MT19937) at 0x1A98C33678, two_step=True,
                verbose=2)
```

[103]:

```
X_filtered = boruta_feature_selector.transform(X)
X_filtered.shape
```

[103]:

```
(642, 280)
```

[104]:

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

[104]:

```
['LASP1',
 'KDM1A',
 'CX3CL1',
 'RHBDF1',
 'PSMB1',
 'MRC2',
 'PTBP1',
 'TMEM159',
 'FHL1',
 'NUP160',
 'SKIV2L2',
 'STAU2',
 'ZIC2',
 'GOPC',
 'R3HDM1',
 'MRTO4',
 'NOP58',
 'ZNF280C',
 'CTSA',
 'WDR18',
 'ERBB3',
 'TMEM206',
 'DIP2B',
 'ZNRD1',
 'KIF2A',
 'NUCKS1',
 'TESK2',
 'PDCD2',
 'NDE1',
 'SCARB1',
 'MARK3',
 'FMO4',
 'ANKRD13A',
 'PAG1',
 'TYR',
 'TP53INP2',
 'DUSP12',
 'CD82',
 'BCORL1',
 'SEH1L',
 'DIMT1',
 'TFAP2C',
 'RFX2',
 'KHSRP',
 'C20orf26',
 'TEKT2',
 'CDC5L',
 'CDC7',
 'HNRNPM',
 'PACSIN2',
 'PRMT5',
 'CEP128',
 'KIAA0247',
 'ZMYND8',
 'ARFGAP1',
 'EEA1',
 'MEDAG',
 'ZNF423',
 'USP31',
 'PIH1D1',
 'SF3A2',
 'ISYNA1',
 'TMEM59L',
 'WDR91',
 'COBL',
 'FUBP3',
 'TRDMT1',
 'NPM3',
 'CUEDC2',
 'SLC6A4',
 'MANBA',
 'GAR1',
 'CRYAB',
 'CPT1A',
 'RNGTT',
 'FANCE',
 'RNF8',
 'BAG2',
 'KHDRBS2',
 'E2F3',
 'WASF1',
 'VNN2',
 'MDFI',
 'BYSL',
 'GHR',
 'TCERG1',
 'NCL',
 'ELMOD3',
 'ORC2',
 'SUMO1',
 'FARSB',
 'EBNA1BP2',
 'CD3EAP',
 'CASP8AP2',
 'UBE3D',
 'HEATR1',
 'NEK6',
 'IFIT2',
 'RNF2',
 'KHDRBS1',
 'UBL3',
 'NUDT10',
 'NLN',
 'ITIH5',
 'RAB9A',
 'TTPAL',
 'ARFGEF2',
 'ZNFX1',
 'NAGK',
 'SNRPC',
 'MED20',
 'KLHDC3',
 'RIOK1',
 'RPP40',
 'UBA2',
 'PRMT1',
 'FLRT1',
 'DNAJC8',
 'IFI6',
 'KRI1',
 'X.13',
 'PHF10',
 'LSM7',
 'EXOSC2',
 'ZNF227',
 'DHX30',
 'FCRLA',
 'MYBBP1A',
 'COQ3',
 'PPP1R3D',
 'CCNA1',
 'RNF128',
 'ANKRD32',
 'WDR74',
 'LRRIQ1',
 'LARS',
 'ELP3',
 'CAPRIN1',
 'CD63',
 'CCT7',
 'ACBD6',
 'ITM2C',
 'DNAJB2',
 'ZFP37',
 'PPIL1',
 'TGS1',
 'TMPRSS13',
 'LRRC49',
 'ACTR1A',
 'BBS7',
 'FBN2',
 'GUCD1',
 'RDH16',
 'SERPINA10',
 'IGF1R',
 'SH3GL3',
 'PARN',
 'COPS3',
 'RPL11',
 'KIAA0319L',
 'ITGB3BP',
 'IGSF3',
 'UCK2',
 'PFDN2',
 'XPR1',
 'SETDB1',
 'LBR',
 'SYT2',
 'ALDH1L1',
 'NCEH1',
 'DGKQ',
 'CAMK2D',
 'RPS3A',
 'RPL7L1',
 'IRAK1BP1',
 'MMS22L',
 'PM20D2',
 'RBMX',
 'GSN',
 'RPP30',
 'FRA10AC1',
 'PPRC1',
 'PDCD11',
 'ENDOD1',
 'FADS1',
 'LIX1L',
 'HOMER1',
 'CWC27',
 'TXNDC11',
 'CLGN',
 'SYCP2L',
 'FBXO36',
 'BUB3',
 'PHKG2',
 'DHRS1',
 'TAB3',
 'SKI',
 'GDPD5',
 'GART',
 'PSMD4',
 'AMFR',
 'ZNF222',
 'CBS',
 'U2AF1',
 'PKN3',
 'DUSP14',
 'NUP35',
 'CCDC138',
 'HDAC11',
 'CADPS',
 'UVSSA',
 'INTU',
 'ABCE1',
 'GRPEL2',
 'DNAAF2',
 'OTX2',
 'PDZD8',
 'DDX21',
 'NOLC1',
 'CCT2',
 'C11orf74',
 'TMED3',
 'CENPV',
 'BLCAP',
 'FAM102A',
 'KIAA1586',
 'MLKL',
 'GJB1',
 'RALGAPB',
 'PA2G4',
 'FAM98B',
 'POLR1C',
 'ANO5',
 'PWWP2B',
 'FRMD5',
 'MAL',
 'MANEA',
 'PURG',
 'CCDC41',
 'LRFN4',
 'OR2T1',
 'PLEKHF2',
 'PFAS',
 'SERTAD2',
 'PDXDC1',
 'CCDC149',
 'GLUD2',
 'RPL35A',
 'CNOT10',
 'SLC25A21',
 'RUVBL2',
 'UTP11L',
 'TANGO2',
 'TMEM106A',
 'NLRP9',
 'ARHGAP30',
 'RDM1',
 'TRMT2B',
 'TMEM120A',
 'TDRD7',
 'SUPT3H',
 'IARS',
 'NTNG2',
 'PTPN1',
 'MYL6B',
 'HDAC2',
 'SLC39A10',
 'SVIL',
  'GSTK1',
  'CXorf40A',
  'SLC9A8',
  'MAK16',
  'X05.Mar',
  'TMEM229B',
  'TMA16',
  'FAM169A',
  'FAM5B',
  'BHLHB9',
  'LIPN',
  'PFDN6']
```

[105]:

```
s_feats = pd.DataFrame(final_features)
s_feats.to_csv('cleaned/boruta-99-25-0.01.csv', index=False)
```

[ ]:

```

```