Automated Source Classification with Random Forests

or Star Bars: A New Hope

Abhigna Revuru

Remeis Observatory & IIT Kharagpur, India

10 July 2018

What is this about?

- Aim - Automate source classification for *near-real time* satellite data, specifically for eROSITA
- Start with - data from 130k sources (hardness ratios, multiwavelength data, spectral features, variability data...)
- Output - what is that thing in the sky? (star, galaxy, black hole, AGN, binary...)

eROSITA

extended ROentgen Survey with an Imaging Telescope Array

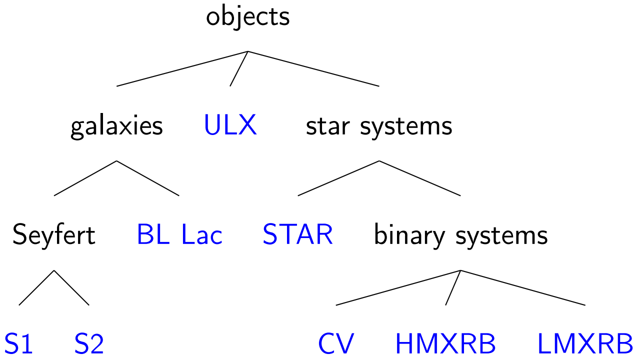
- primary instrument aboard the Russian SRG mission
- expected launch: Baikonur, March 2019
- deep survey of x-ray sky in 0.5-10keV band

eROSITA mirror modules

Technical Specifications

Energy range	0.2 - 10keV
Energy resolution	138eV at 6keV
Focal length	1.6m
Field of view	61 arcsec
Effective area	1400 sqcm at 1keV
Time resolution	50ms

Source classes



Galaxies

- luminous nuclei
- high surface brightness
- high ionisation emission line spectra

Seyfert I galaxies

- Broad H lines
- Narrow forbidden lines from H, He, O

Seyfert II galaxies

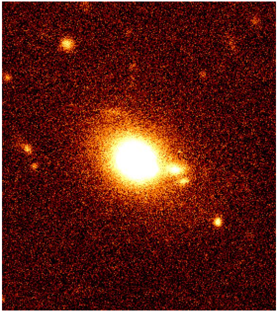
- No broad emission lines
- Strong absorption

NGC 1068, one of the first Seyfert galaxies classified

BL Lac Objects

galaxies with AGNs, named after its prototype, BL Lacertae

- rapid and large-amplitude flux variability
- relatively featureless spectra



H 0323+022, a BL Lac object, with visible host galaxy and close companions

X-Ray Binaries

stellar ***donor*** + compact, black hole/neutron star ***accretor***

High Mass X-ray Binaries (HMXRB)

- donor: massive star/blue supergiant
- mass transfer via donor's stellar wind captured by accretor

Low Mass X-ray Binaries (LMXRB)

- donor: main sequence star/white swarf/red giant
- mass transfer from donor Roche lobe to accretor

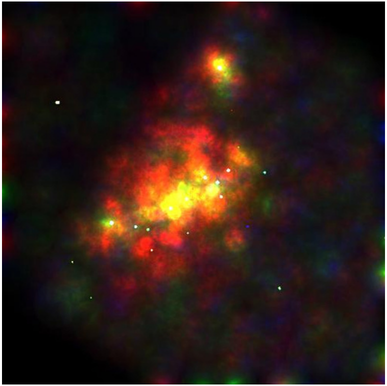
Cataclysmic Variable Stars

accreting white dwarf + mass transferring ***donor*** star

- binary stars
- irregular increase in brightness by a large factor, then drops down to inactive state
- easy to classify - rapid variability, luminous, peculiar emission lines

Ultraluminous x-ray sources

- less luminous than AGNs
- but more luminous than any other known stellar process
- luminosity exceeds that of neutron stars and stellar black holes



Chandra image of NGC 4485 and NGC 4490: two potential ULXs

Using machine learning

How to classify an unknown source?

Traditional way

- crossmatch source positions with catalogues of other wavelengths
- workflow: spectral fitting and intuitive classification rules
- consumes human effort, time

With ML

- each 'feature' of the data is mapped by an unknown function to the source class
- workflow: optimizing this unknown function for highest accuracy
- reproducible, efficient, scalable

Examples of machine learning

1. Supervised
   - me telling you the difference
   - 'teach' the algorithm what conclusions it should come up with
   - eg: predicting world cup outcomes based on old football statistics
2. Unsupervised
   - you figuring out the difference for yourself
   - let the algorithm look for patterns in data by itself
   - eg: distinguishing between pictures of chairs, aeroplanes and unicorns

What is this black magic?

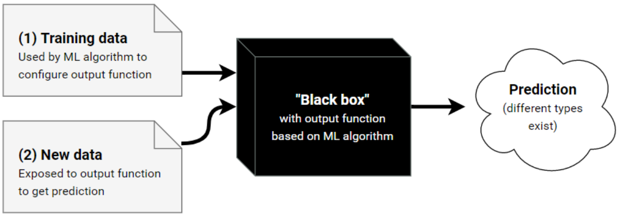
Goal: to generalise beyond the training set

$$\mathbf{y} = \mathbf{F}(\mathbf{a},\mathbf{b},\mathbf{c},\ldots)$$

- **y** is a response variable
- **a,b,c,...** are features of the data
- What is **F**?

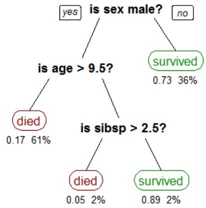
  → don't know, don't care

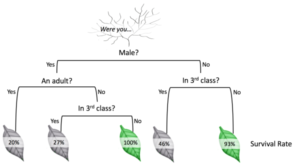
  → our job to find a blackbox that performs the best

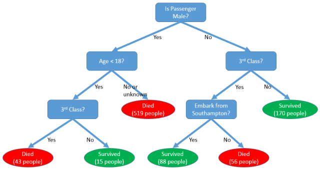


Blackbox example 1

Survival on the Titanic - predictor/classifier



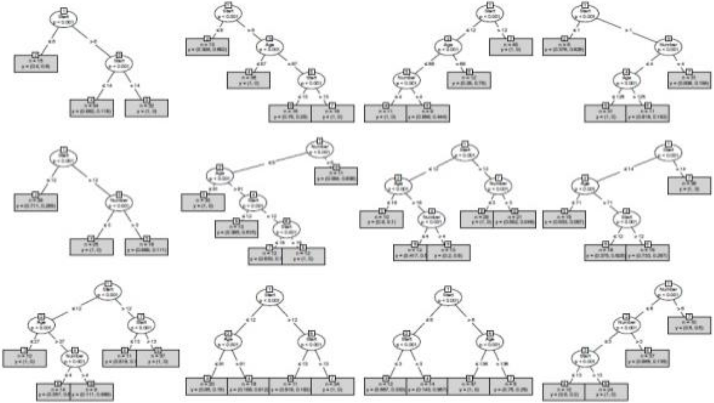




"decision tree"

Blackbox example 2

What if we use multiple instances of a decision tree?
$\Rightarrow$ It becomes a forest!



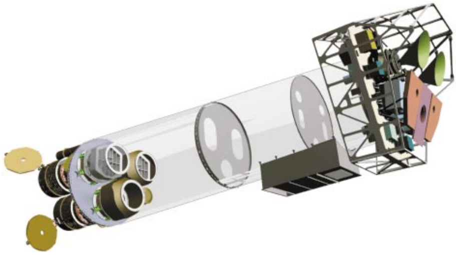
Random forest algorithm

- one tree = decisions get distorted by sparse training data after a while $\Rightarrow$ overfitting
- multiple instances reduce noise and average out errors
- an *ensemble* = complementary trees that contribute to a single effect
- 'random' $\rightarrow$ each tree is a randomised sample/subset of the training set (*bagging*)
  $\Rightarrow$ Unique trees, so different classifications, votes are tallied at the end
- Classifier parameters - number of trees, minimum split, split criteria, etc

Implementing random forests

Case in point: the EXTraS project

- Maintainer: ESA
- Launch: December 1999
- Focal length: 7.4 m
- Range: 0.15 keV – 12 keV
- Resolution: 150 eV

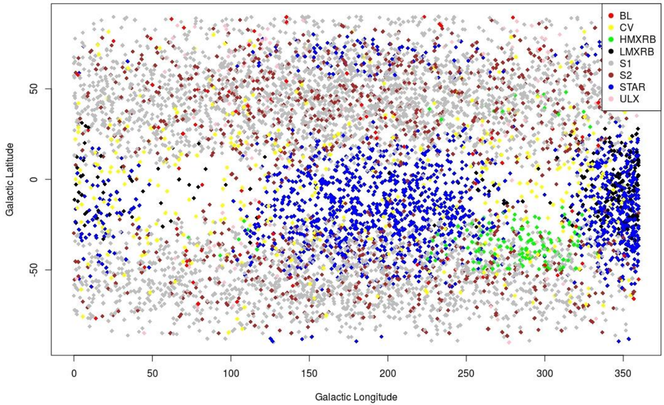


XMM-Newton (Jansen et al., 2001)

Exploring the **X**-ray and **Tra**nsient Variable **S**ky
to investigate unexplored archival data from the cameras onboard XMM-Newton

The training dataset

- Requirement: must be representative of properties of various sources to ensure a faithful classification
- **Farrell et al 2015** - an early attempt at using random forests
- 7383 detections of 2911 sources over 8 source classes



Balancing the training data

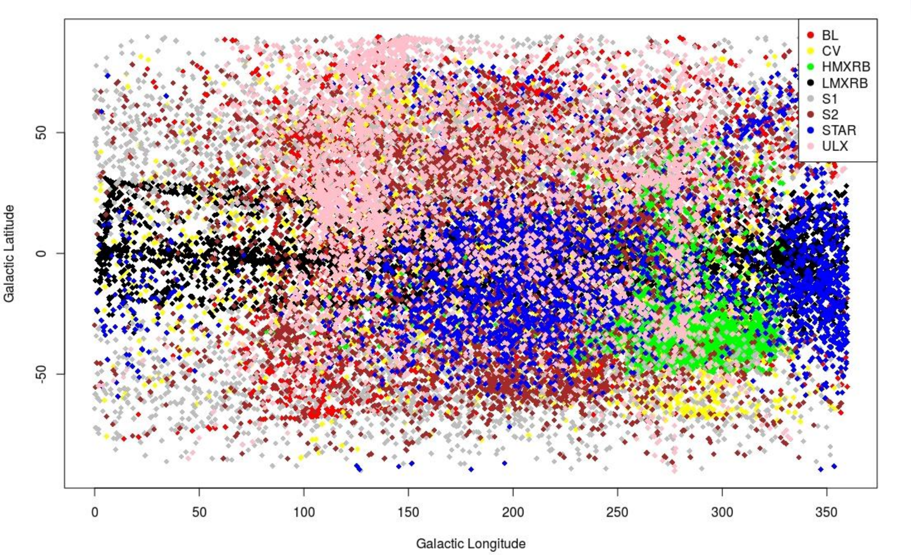
- Data set is heavily skewed towards galaxies
- SMOTE algorithm - 'Synthetic minority over-sampling technique' from Chawla et al 2011
- oversampling minority classes + undersampling majority class

Source type	Number of sources		Number of detections	
BL	60	(2%)	104	(2%)
CV	201	(7%)	396	(5.5%)
HMXRB	33	(1%)	150	(2.5%)
LMXRB	66	(2.5%)	206	(3%)
STAR	563	(19%)	1613	(21%)
Seyfert 1 (S1)	1486	(51%)	3773	(51%)
Seyfert 2 (S2)	485	(17%)	1026	(13%)
ULX	17	(0.5%)	115	(2%)
Total	**2911**		**7383**	

Farrell set

Source type	Oversampling	Number of detections
BL	3300	3536
CV	800	3564
HMXRB	1600	3750
LMXRB	2400	3502
STAR	150	3226
Seyfert 1 (S1)	-	3773
Seyfert 2 (S2)	300	4104
ULX	3300	3910
Total	**-**	**20 365**

SMOTEd set



29365 observations in the balanced training set with a coordinate bias

Parameter optimization

```
library(randomForest)
fit <- randomForest(class~., data=train_par, importance=T, mtry=6, ntree=600, na.action=na.roughfix)
```

- Train on parameter **class**
- Assesing importance of predictors
- **ntree** - trees grown
- **mtry** - variables randomly sampled as candidates at each split, close to usually sqrt(number of variables)
- can optimise by hit/trial or with *caret* package

Feature selection

232 features available for every detection - which ones are important?

- Object position
- Hardness ratios
- Spectral parameters from 6 models (powerlaw, blackbody, ionised plasma, power law + instrinsically absorbed power law,..)
- Multiwavelength indices
- Timing - signal ffr, probability, power, exposure, count rates, fractional variability

Hardness Ratio	lower band [keV]	upper band [keV]
HR1	0.2–0.5	0.5–1.0
HR2	0.5–1.0	1.0–2.0
HR3	1.0–2.0	2.0–4.5
HR4	2.0–4.5	4.5–12.0

Results

coordinates

- accuracy = 60%
- variable importance:

```
> importance(fit, type=1)
              MeanDecreaseAccuracy
threexmm_bii              797.7606
threexmm_lii              686.5292
> importance(fit, type=2)
              MeanDecreaseGini
threexmm_bii          13042.26
threexmm_lii          12622.47
```

```
               Type of random forest: classification
                     Number of trees: 600
No. of variables tried at each split: 2

        OOB estimate of  error rate: 59.4%
Confusion matrix:
        BL  CV HMXRB LMXRB   S1   S2 STAR  ULX class.error
BL    1154 237   141   244  481  551  259  469   0.6736425
CV     258 857   415   509  304  434  399  388   0.7595398
HMXRB   77 231  2843   126   59  130  137  147   0.2418667
LMXRB  140 400   136  1868   57  193  502  206   0.4665905
S1     457 241   136    96 1480  682  205  476   0.6077392
S2     526 422   191   250  756 1066  416  477   0.7402534
STAR   250 341   196   556  199  362 1116  206   0.6540608
ULX    406 337   263   223  479  456  208 1538   0.6066496
```

hardness ratios

- accuracy = 91.4%
- variable importance:

```
> importance(fit, type = 1)
            MeanDecreaseAccuracy
cat3xmm_hr1            1656.8226
cat3xmm_hr2             513.2138
cat3xmm_hr3            1397.8394
cat3xmm_hr4             926.0396
> importance(fit, type = 2)
            MeanDecreaseGini
cat3xmm_hr1         8425.331
cat3xmm_hr2         4910.605
cat3xmm_hr3         6432.381
cat3xmm_hr4         5907.504
```

```
               Type of random forest: classification
                     Number of trees: 600
No. of variables tried at each split: 4

        OOB estimate of  error rate: 8.63%
Confusion matrix:
        BL   CV HMXRB LMXRB   S1   S2 STAR  ULX class.error
BL    3372   24     0    11   67   25    9   28  0.04638009
CV      93 2907    32    78  189  218   22   25  0.18434343
HMXRB    0   11  3707    12    0   20    0    0  0.01146667
LMXRB    2   21    14  3345    8   26   30   56  0.04483152
S1      77  153     7    23 3296  177   22   18  0.12642460
S2      38  195    33   104  221 3428   47   38  0.16471735
STAR    14    8     4    85   17   38 2974   86  0.07811531
ULX     12    7     0    30    9   15   34 3803  0.02736573
```

coordinates + HR

- accuracy = 92.5%
- variable importance:

```
> importance(fit, type = 1)
             MeanDecreaseAccuracy
threexmm_lii             327.5280
threexmm_bii             388.9156
cat3xmm_hr1             1005.3767
cat3xmm_hr2              290.2397
cat3xmm_hr3              623.0622
cat3xmm_hr4              644.6072
> importance(fit, type = 2)
             MeanDecreaseGini
threexmm_lii         1842.123
threexmm_bii         3133.720
cat3xmm_hr1          6925.016
cat3xmm_hr2          3849.591
cat3xmm_hr3          5372.199
cat3xmm_hr4          4552.909
```

```
               Type of random forest: classification
                     Number of trees: 600
No. of variables tried at each split: 5

        OOB estimate of  error rate: 7.56%
Confusion matrix:
        BL   CV HMXRB LMXRB   S1   S2 STAR  ULX class.error
BL    3402   24     0     4   55   17    8   26 0.037895928
CV      80 3097    22    49   99  170   24   23 0.131032548
HMXRB    0    2  3737     4    0    6    1    0 0.003466667
LMXRB    1   17     4  3385    4    7   27   57 0.033409480
S1      95  132     4    12 3237  246   20   27 0.142062020
S2      32  159    20    48  269 3463   59   54 0.156189084
STAR    10    7     2    66   12   45 3014   70 0.065716057
ULX      9    5     0    22    7   22   35 3810 0.025575448
```

spectral fit data

- accuracy = 93.8%
- variable importance:

```
> importance(fit, type = 1)
               MeanDecreaseAccuracy
pl_norm                    66.78330
pl_tbnewnh                144.07583
pl_gamma                   54.93509
bb_temp                    62.53436
bb_norm                    82.63826
pl_norm.1                  67.11704
pl_gamma.1                 53.57075
ap_tbnewnh                118.12447
bb_temp.1                  59.19234
ap_temp                    56.47112
bbpl_bbtemp               105.60490
bbpl_gamma                 78.73861
appl_redstat               77.15971
bb_redstat                 82.30634
bbpl_redstat               83.98089
pl_redstat                125.52846
plpl_redstat              107.64550
ap_redstat                 90.87744
```

```
               Type of random forest: classification
                     Number of trees: 600
No. of variables tried at each split: 5

        OOB estimate of  error rate: 6.28%
Confusion matrix:
```

	BL	CV	HMXRB	LMXRB	S1	S2	STAR	ULX	class.error
BL	3492	3	0	0	24	1	8	8	0.01244344
CV	3	3258	40	5	98	89	53	18	0.08585859
HMXRB	0	1	3710	0	1	36	2	0	0.01066667
LMXRB	2	7	1	3429	10	10	24	19	0.02084523
S1	29	57	31	29	3331	232	51	13	0.11714816
S2	7	57	78	13	427	3426	75	21	0.16520468
STAR	6	27	6	3	51	78	3023	32	0.06292622
ULX	6	8	0	0	22	2	20	3852	0.01483376

coordinates + HR + timing

- accuracy = 95.2%
- variable importance:

```
> importance(fit, type = 1)
              MeanDecreaseAccuracy
threexmm_bii             223.55524
threexmm_lii             146.65100
cat3xmm_hr1              465.50555
cat3xmm_hr2              169.69160
cat3xmm_hr3              322.56389
cat3xmm_hr4              402.72402
ffr                      114.62382
probability               92.54753
power                     78.29887
cat3xmm_fvar             326.98034
> importance(fit, type = 2)
              MeanDecreaseGini
threexmm_bii         2235.8360
threexmm_lii         1343.5020
cat3xmm_hr1          5747.8457
cat3xmm_hr2          3095.1287
cat3xmm_hr3          4288.2215
cat3xmm_hr4          3809.9307
ffr                   707.6687
probability           667.6979
power                 491.4219
cat3xmm_fvar         3288.7555
```

```
               Type of random forest: classification
                     Number of trees: 600
No. of variables tried at each split: 5

        OOB estimate of  error rate: 4.78%
Confusion matrix:
        BL   CV HMXRB LMXRB   S1   S2 STAR  ULX class.error
BL    3479    7     0     3   31    7    1    8 0.016119910
CV      47 3276    10    14   47  126   26   18 0.080808081
HMXRB    0    1  3746     1    0    2    0    0 0.001066667
LMXRB    0    6     0  3440    2   15   25   14 0.017704169
S1      63   54     4    16 3388  212   12   24 0.102040816
S2      31   81    13    56  241 3640   29   13 0.113060429
STAR     7   13     0    19    2   24 3120   41 0.032858029
ULX      2    6     0     2    1    4   23 3872 0.009718670
```

coordinates + HR + MWL

- accuracy = 95.7%
- variable importance:

```
> importance(fit, type = 1)
               MeanDecreaseAccuracy
threexmm_lii               219.9314
threexmm_bii               250.5761
cat3xmm_hr1                750.8883
cat3xmm_hr2                222.7968
cat3xmm_hr3                422.8834
cat3xmm_hr4                358.9138
a_ir1x                     192.2521
a_r1x                      524.4281
a_ox                       155.2701
a_gx                       296.6534
> importance(fit, type = 2)
               MeanDecreaseGini
threexmm_lii           882.8117
threexmm_bii          1723.0129
cat3xmm_hr1           5443.2164
cat3xmm_hr2           2240.9029
cat3xmm_hr3           3979.0328
cat3xmm_hr4           3560.1715
a_ir1x                1769.0224
a_r1x                 2842.7349
a_ox                  1388.9250
a_gx                  1845.9799
```

```
               Type of random forest: classification
                     Number of trees: 600
No. of variables tried at each split: 9

        OOB estimate of  error rate: 4.29%
Confusion matrix:
        BL   CV HMXRB LMXRB   S1   S2 STAR  ULX class.error
BL    3486    4     0     1   24   17    1    3 0.014140271
CV      13 3359    19    39   59   32   29   14 0.057519641
HMXRB    0    2  3740     5    0    0    3    0 0.002666667
LMXRB    0   18     0  3397    4    8   19   56 0.029982867
S1      28   86     1     6 3458  158   22   14 0.083487941
S2      29   67     8    31  209 3702   28   30 0.097953216
STAR    10   28     2    27    7   21 3092   39 0.041537508
ULX      1    2     0    17    4    1   13 3872 0.009718670
```

coordinates + HR + spectral fit

- accuracy = 96.4%
- variable importance:

```
> importance(fit, type = 1)
                 MeanDecreaseAccuracy
threexmm_lii                119.97568
threexmm_bii                160.59532
cat3xmm_hr1                 148.80289
cat3xmm_hr2                  63.37323
cat3xmm_hr3                  80.09766
cat3xmm_hr4                 122.06154
pl_norm                      51.01305
pl_tbnewnh                   87.06941
pl_gamma                     38.90890
bb_temp                      44.05168
bb_norm                      78.41238
pl_norm.1                    49.94937
pl_gamma.1                   38.71991
ap_tbnewnh                   84.64302
bb_temp.1                    42.75562
ap_temp                      44.94677
bbpl_bbtemp                  77.39873
bbpl_gamma                   54.79677
appl_redstat                 54.45982
bb_redstat                   66.26588
bbpl_redstat                 60.75254
pl_redstat                   97.48770
plpl_redstat                 87.20605
ap_redstat                   78.77074
```

```
               Type of random forest: classification
                     Number of trees: 600
No. of variables tried at each split: 9

        OOB estimate of  error rate: 3.56%
Confusion matrix:
        BL   CV HMXRB LMXRB   S1   S2 STAR  ULX class.error
BL    3511    0     0     0   19    1    3    2 0.007070136
CV       5 3379    11     6   44   73   34   12 0.051907969
HMXRB    0    0  3746     0    0    4    0    0 0.001066667
LMXRB    0    4     0  3447    6   10   23   12 0.015705311
S1      28   45     1     9 3452  203   25   10 0.085078187
S2       5   40     9     7  271 3727   31   14 0.091861598
STAR     7    6     0     6    6   30 3159   12 0.020768754
ULX      1    1     0     0    3    1    6 3898 0.003069054
```

coordinates + HR + Spec + timing

- accuracy = 97%
- variable importance:

```
> importance(fit, type = 1)
              MeanDecreaseAccuracy
threexmm_bii              95.64525
threexmm_lii              85.21602
cat3xmm_hr1              102.34006
cat3xmm_hr2               52.49468
cat3xmm_hr3               67.64200
cat3xmm_hr4               91.13762
ffr                       63.26643
probability               60.08891
power                     54.08407
cat3xmm_fvar             117.68353
pl_norm                   55.82307
pl_tbnewnh                66.02866
pl_gamma                  40.79278
bb_temp                   39.99804
bb_norm                   63.97834
pl_norm.1                 53.41880
pl_gamma.1                38.99407
ap_tbnewnh                57.08867
bb_temp.1                 37.47801
ap_temp                   35.51197
bbpl_bbtemp               56.37222
bbpl_gamma                44.54577
appl_redstat              41.74002
bb_redstat                52.78453
bbpl_redstat              52.19096
pl_redstat                72.23750
plpl_redstat              72.47268
ap_redstat                57.42948
```

```
               Type of random forest: classification
                     Number of trees: 600
No. of variables tried at each split: 6

        OOB estimate of  error rate: 3.09%
Confusion matrix:
        BL   CV HMXRB LMXRB   S1   S2 STAR  ULX  class.error
BL    3516    0     0     0   15    1    2    2 0.0056561086
CV       2 3403     8     5   40   64   32   10 0.0451739618
HMXRB    0    0  3748     0    0    1    1    0 0.0005333333
LMXRB    0    5     0  3456    5    7   20    9 0.0131353512
S1      26   26     1     8 3504  183   16    9 0.0712960509
S2       7   19     5     7  265 3764   26   11 0.0828460039
STAR     7    7     0     4    7   24 3166   11 0.0185988841
ULX      0    1     0     0    2    1    4 3902 0.0020460358
```

Optimum features to train with

```
> importance(fit, type = 1)
               MeanDecreaseAccuracy
threexmm_bii               84.26848
threexmm_lii               71.87563
cat3xmm_hr1               183.31371
cat3xmm_hr2                67.09111
cat3xmm_hr3               103.33551
cat3xmm_hr4               111.28152
cat3xmm_e_hr1              64.30636
cat3xmm_e_hr2              39.50628
cat3xmm_e_hr3              41.94883
cat3xmm_e_hr4              52.25738
cat3xmm_flux8             127.76902
cat3xmm_fvar              125.24421
cat3xmm_s                 110.05040
cat3xmm_c                   0.00000
a_ir1x                     73.67552
a_r1x                     147.15926
a_ox                       71.52170
a_gx                      108.20562
ffr                        51.23457
probability                59.60276
power                      57.69735
```

```
         Type of random forest: classification
               Number of trees: 600
No. of variables tried at each split: 7

        OOB estimate of  error rate: 1.6%
Confusion matrix:
        BL   CV HMXRB LMXRB   S1   S2 STAR  ULX  class.error
BL    3526    0     0     0    8    2    0    0 0.0028280543
CV       1 3507     1     5   29    4   17    0 0.0159932660
HMXRB    0    0  3750     0    0    0    0    0 0.0000000000
LMXRB    0    4     0  3494    0    0    4    0 0.0022844089
S1       8   21     0     4 3604  119   14    3 0.0447919428
S2       4   10     3     2  162 3903   17    3 0.0489766082
STAR     2    3     0     1    6    7 3202    5 0.0074395536
ULX      0    0     0     0    0    0    1 3909 0.0002557545
```

Using 20 features, accuracy = 98.4%

Adding spectral fit data

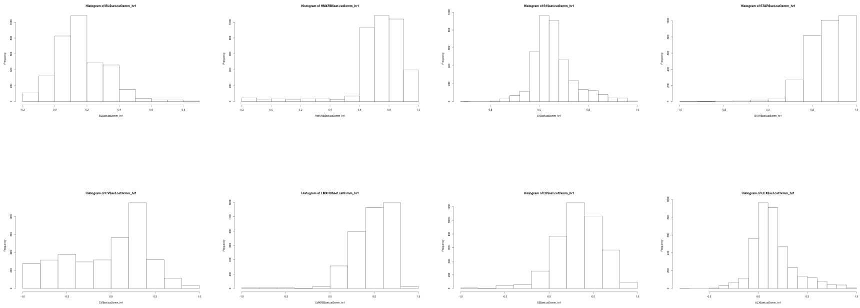
```
               Type of random forest: classification
                     Number of trees: 600
No. of variables tried at each split: 6

        OOB estimate of  error rate: 1.62%
Confusion matrix:
```

	BL	CV	HMXRB	LMXRB	S1	S2	STAR	ULX	class.error
BL	3530	0	0	0	5	1	0	0	0.0016968326
CV	0	3513	1	3	26	2	19	0	0.0143097643
HMXRB	0	0	3750	0	0	0	0	0	0.0000000000
LMXRB	0	5	0	3491	0	0	6	0	0.0031410623
S1	4	12	0	9	3602	129	16	1	0.0453220249
S2	6	7	0	2	181	3893	12	3	0.0514132554
STAR	0	6	0	2	6	8	3201	3	0.0077495350
ULX	0	0	0	0	0	0	1	3909	0.0002557545

When we include 6 spectral features that contribute the most, accuracy = 98.38%

Why this indifference?

Hardness ratio dependence



cat3xmm_hr1 - Histograms of BL, HMXRB, S1, STAR, ULX, S2, LMXRB, CV (in clockwise order)

More algorithms to look into

- Algorithms - k-nearest neighbour, support vector machines, neural networks
- Evaluation - precision/recall sensitivity, cost/utility margin
- Optimization

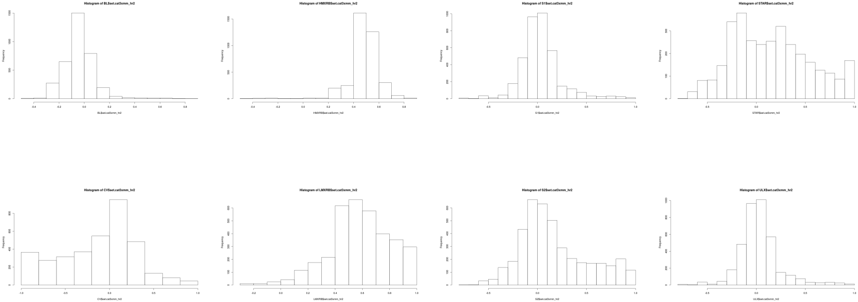
Representation	Evaluation	Optimization
Instances	Accuracy/Error rate	Combinatorial optimization
K-nearest neighbor	Precision and recall	Greedy search
Support vector machines	Squared error	Beam search
Hyperplanes	Likelihood	Branch-and-bound
Naive Bayes	Posterior probability	Continuous optimization
Logistic regression	Information gain	Unconstrained
Decision trees	K-L divergence	Gradient descent
Sets of rules	Cost/Utility	Conjugate gradient
Propositional rules	Margin	Quasi-Newton methods
Logic programs		Constrained
Neural networks		Linear programming
Graphical models		Quadratic programming
Bayesian networks		
Conditional random fields		

Final thoughts

- Can we algorithmise the way we think about sources?
- Can we teach a machine to understand more complex spectral models?
- Transiting from classifying to clustering to find new associations

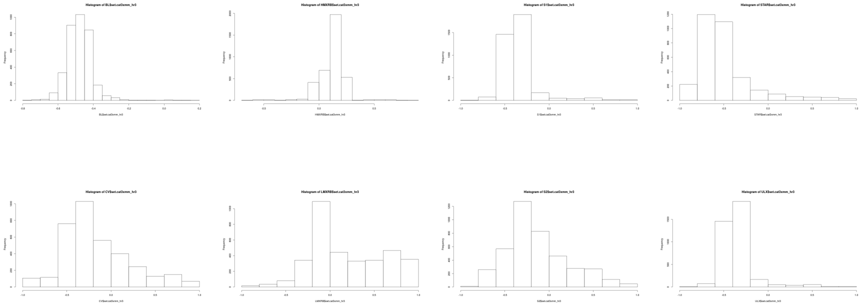
Questions?

Hardness ratio dependence (2)



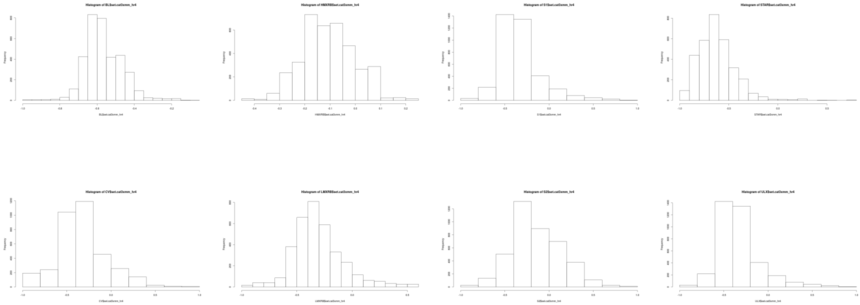
cat3xmm_hr2 - Histograms of BL, HMXRB, S1, STAR, ULX, S2, LMXRB, CV (in clockwise order)

Hardness ratio dependence (3)



cat3xmm_hr3 - Histograms of BL, HMXRB, S1, STAR, ULX, S2, LMXRB, CV (in clockwise order)

Hardness ratio dependence (4)



cat3xmm_hr4 - Histograms of BL, HMXRB, S1, STAR, ULX, S2, LMXRB, CV (in clockwise order)