Implementation of Random Forest

Christian Reiser

Algorithm

`ID3` is used as a base learner. `ID3` works by considering at each split every attribute, that has not already been used on a split upwards in the tree.. For that matter we calculate the subsets that would result from splitting on a specific attribute. For each attribute value a different subset is generated. The entropy of each subset can then be analyzed. We favor subsets with low entropy, i.e. subsets that are as homogenous as possible. The entropies of the subsets are then summarized in the information gain value by relating the size of the subsets with their entropy. A big subset has a higher impact on the information gain value. The purpose of the information gain value is to approximate how beneficial a split on a specific attribute would be. Therefore the split with the highest information gain is selected at each node.
Additionally `Bagging` is used: instead of generating only one classifier (in this case a decision tree) we create multiple ones. The idea is to build classifiers that are as uncorrelated as possible. A strategy to achieve that is to train each classifier on a different training set. In the case of `Bagging` we build the new training sets from the original training set by randomly taking `n` samples from the training set, where `n` is the size of the training set. By that we get a training set with some duplicate samples and some missing samples. Duplicate samples make the associated classifier prioritize the correct classification of that sample. To get a final prediction of the ensemble of trees Majority Voting is used.
On top of `Bagging` we use the `Random Forest` technique. At each split only a random subset of a fixed size of attributes is considered for splitting. By that the resulting trees are even further decorrelated, since the classifier is forced to use other attributes for splitting.

Execution of the code

Python with `numpy` and `pandas` packages installed is required. To execute the algorithm the filename of the dataset must be specified as a command line argument. Since the column that contains the class label is different from dataset to dataset this can be specified as an additional command line argument. To reproduce the results from this report the following three commands should be executed:

```
python ./randomforest.py ../Data/house-votes-84.data -t 0.9 -c first
python ./randomforest.py ../Data/agaricus-lepiota.data -t 0.9 -c first
python ./randomforest.py ../Data/car.data -t 0.9 -c last
```

The `-t` parameter specifies the percentage of the dataset that is used for testing, while `-c` specifies the column that contains the class label. With parameter `-nt` the number of trees can be specified. The number of features considered at each split can be controlled via the parameter `-fss`. For more detailed information you can call the help function:

```
python ./rules -h
```

Evaluation

Three datasets from UCI ML were used

- http://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records
- http://archive.ics.uci.edu/ml/datasets/Car+Evaluation
- http://archive.ics.uci.edu/ml/datasets/Mushroom

For all datasets only 10% of the examples were used for training, so the remaining 90% could be used for testing.

In the following table the number of trees are varied over the rows ($NT \in \{50,\ 100\}$) and the number of features are varied over the columns ($F \in \{1,\ 3,\ int(log_2(M)),\ \sqrt{M}\ \}$). In each inner cell the first value corresponds to the Congressional Voting Records dataset, the second values corresponds to the Car Evaluation dataset and the third value corresponds to Mushroom dataset.

	1	3	$int(log_2(M))$	$\sqrt{M}$
50	91.84% 99.88% 83.99%	94.90% 100.00% 85.79%	95.41% 99.89% 86.11%	94.39% 99.89% 85.72%
100	91.07% 99.97% 82.83%	94.39% 99.89% 85.66%	95.15% 99.96% 86.24%	93.37% 99.89% 86.05%

The feature importance is calculated as the average information gain at each split the corresponding feature was used.

Congressional Voting Records

	$NT = 50,\ F = 1$	
1	physician-fee-freeze	0.515
2	water-project-cost-sharing	0.306
3	duty-free-exports	0.286
4	adoption-of-the-budget-resolution	0.264
5	superfund-right-to-sue	0.213
6	synfuels-corporation-cutback	0.193
7	immigration	0.19
8	el-salvador-aid	0.182
9	education-spending	0.174
10	aid-to-nicaraguan-contras	0.164
11	mx-missile	0.162

12	export-administration-act-south-africa	0.157
13	handicapped-infants	0.146
14	crime	0.136
15	anti-satellite-test-ban	0.134
16	religious-groups-in-schools	0.129
	$NT = 50,\ F = 3$	
1	physician-fee-freeze	0.437
2	water-project-cost-sharing	0.403
3	export-administration-act-south-africa	0.368
4	duty-free-exports	0.366
5	adoption-of-the-budget-resolution	0.341
6	mx-missile	0.299
7	superfund-right-to-sue	0.286
8	immigration	0.285
9	education-spending	0.28
10	synfuels-corporation-cutback	0.275
11	handicapped-infants	0.256
12	el-salvador-aid	0.244
13	anti-satellite-test-ban	0.21
14	religious-groups-in-schools	0.192
15	crime	0.157
16	aid-to-nicaraguan-contras	0.149
	$NT = 50,\ F = int(log_2(M))$	
1	physician-fee-freeze	0.422
2	aid-to-nicaraguan-contras	0.396
3	water-project-cost-sharing	0.381
4	export-administration-act-south-africa	0.368
5	immigration	0.351
6	adoption-of-the-budget-resolution	0.347
7	education-spending	0.337
8	synfuels-corporation-cutback	0.317
9	el-salvador-aid	0.316
10	duty-free-exports	0.309

11	handicapped-infants	0.301
12	mx-missile	0.289
13	crime	0.289
14	superfund-right-to-sue	0.242
15	anti-satellite-test-ban	0.199
16	religious-groups-in-schools	0.198
	$NT = 50, F = \sqrt{M}$	
1	water-project-cost-sharing	0.431
2	physician-fee-freeze	0.427
3	duty-free-exports	0.409
4	adoption-of-the-budget-resolution	0.359
5	el-salvador-aid	0.322
6	education-spending	0.311
7	export-administration-act-south-africa	0.308
8	superfund-right-to-sue	0.279
9	synfuels-corporation-cutback	0.276
10	anti-satellite-test-ban	0.262
11	handicapped-infants	0.258
12	mx-missile	0.25
13	immigration	0.229
14	crime	0.218
15	aid-to-nicaraguan-contras	0.208
16	religious-groups-in-schools	0.0821
	$NT = 100, F = 1$	
1	physician-fee-freeze	0.495
2	duty-free-exports	0.344
3	adoption-of-the-budget-resolution	0.251
4	water-project-cost-sharing	0.247
5	immigration	0.229
6	superfund-right-to-sue	0.212
7	export-administration-act-south-africa	0.196
8	crime	0.196
9	anti-satellite-test-ban	0.192
10	synfuels-corporation-cutback	0.182

11	el-salvador-aid	0.176
12	handicapped-infants	0.17
13	mx-missile	0.168
14	aid-to-nicaraguan-contras	0.163
15	education-spending	0.153
16	religious-groups-in-schools	0.147
	$NT = 100,\ F = 3$	
1	physician-fee-freeze	0.452
2	adoption-of-the-budget-resolution	0.338
3	water-project-cost-sharing	0.334
4	immigration	0.323
5	anti-satellite-test-ban	0.319
6	aid-to-nicaraguan-contras	0.3
7	el-salvador-aid	0.292
8	duty-free-exports	0.29
9	export-administration-act-south-africa	0.277
10	mx-missile	0.275
11	synfuels-corporation-cutback	0.261
12	superfund-right-to-sue	0.26
13	religious-groups-in-schools	0.247
14	handicapped-infants	0.243
15	education-spending	0.219
16	crime	0.165
	$NT = 100,\ F = int(log_2(M))$	
1	physician-fee-freeze	0.454
2	water-project-cost-sharing	0.421
3	education-spending	0.354
4	export-administration-act-south-africa	0.327
5	adoption-of-the-budget-resolution	0.322
6	immigration	0.315
7	synfuels-corporation-cutback	0.309
8	duty-free-exports	0.305
9	handicapped-infants	0.297
10	aid-to-nicaraguan-contras	0.295

11	superfund-right-to-sue	0.285
12	el-salvador-aid	0.267
13	crime	0.265
14	mx-missile	0.261
15	anti-satellite-test-ban	0.252
16	religious-groups-in-schools	0.171
	$NT = 100,\ F = \sqrt{M}$	
1	physician-fee-freeze	0.463
2	adoption-of-the-budget-resolution	0.389
3	duty-free-exports	0.363
4	mx-missile	0.357
5	export-administration-act-south-africa	0.35
6	water-project-cost-sharing	0.325
7	superfund-right-to-sue	0.318
8	immigration	0.301
9	el-salvador-aid	0.295
10	synfuels-corporation-cutback	0.291
11	aid-to-nicaraguan-contras	0.281
12	handicapped-infants	0.271
13	education-spending	0.254
14	anti-satellite-test-ban	0.239
15	crime	0.23
16	religious-groups-in-schools	0.161

Mushroom

	$NT = 50,\ F = 1$	
1	odor	0.476
2	spore-print-color	0.361
3	habitat	0.345
4	stalk-root	0.327
5	gill-size	0.304
6	cap-color	0.287
7	gill-color	0.285

8	population	0.262
9	stalk-shape	0.248
10	cap-surface	0.227
11	bruises	0.203
12	ring-type	0.194
13	cap-shape	0.161
14	stalk-surface-below-ring	0.152
15	stalk-color-below-ring	0.145
16	gill-spacing	0.144
17	stalk-surface-above-ring	0.128
18	ring-number	0.107
19	stalk-color-above-ring	0.1
20	gill-attachment	0.025
21	veil-color	0.0102
22	veil-type	0
	$NT = 50,\ F = 3$	
1	odor	0.515
2	bruises	0.444
3	gill-size	0.408
4	stalk-root	0.4
5	spore-print-color	0.39
6	gill-color	0.345
7	habitat	0.341
8	population	0.323
9	ring-type	0.302
10	cap-surface	0.293
11	gill-spacing	0.269
12	cap-color	0.267
13	stalk-surface-below-ring	0.262
14	stalk-shape	0.249
15	cap-shape	0.24
16	stalk-color-below-ring	0.229
17	ring-number	0.211
18	stalk-surface-above-ring	0.206

19	stalk-color-above-ring	0.197
20	gill-attachment	0.104
21	veil-color	0.0489
22	veil-type	0
	$NT = 50,\ F\ =\ int(log_2(M))$	
1	odor	0.505
2	bruises	0.391
3	gill-spacing	0.387
4	stalk-root	0.379
5	gill-size	0.353
6	habitat	0.351
7	ring-type	0.34
8	population	0.334
9	ring-number	0.325
10	spore-print-color	0.324
11	cap-color	0.317
12	stalk-surface-below-ring	0.317
13	stalk-surface-above-ring	0.292
14	cap-surface	0.283
15	stalk-shape	0.271
16	cap-shape	0.266
17	stalk-color-below-ring	0.258
18	gill-color	0.255
19	veil-color	0.241
20	stalk-color-above-ring	0.214
21	gill-attachment	0.121
22	veil-type	0
	$NT = 50,\ F = \sqrt{M}$	
1	odor	0.483
2	stalk-root	0.376
3	gill-size	0.363
4	bruises	0.362
5	spore-print-color	0.361
6	cap-surface	0.345

7	stalk-shape	0.344
8	stalk-surface-above-ring	0.314
9	population	0.31
10	habitat	0.302
11	cap-color	0.298
12	gill-color	0.282
13	stalk-surface-below-ring	0.274
14	ring-type	0.268
15	gill-spacing	0.257
16	stalk-color-below-ring	0.255
17	cap-shape	0.245
18	ring-number	0.233
19	veil-color	0.225
20	stalk-color-above-ring	0.16
21	gill-attachment	0.0333
	$NT = 100,\ F = 1$	
1	odor	0.5
2	spore-print-color	0.351
3	habitat	0.33
4	gill-color	0.315
5	stalk-root	0.311
6	population	0.305
7	cap-color	0.283
8	gill-size	0.272
9	bruises	0.254
10	cap-surface	0.221
11	ring-type	0.205
12	stalk-shape	0.19
13	stalk-surface-below-ring	0.175
14	cap-shape	0.171
15	stalk-color-below-ring	0.153
16	gill-spacing	0.137
17	stalk-surface-above-ring	0.128
18	stalk-color-above-ring	0.123

19	ring-number	0.0968
20	veil-color	0.014
21	gill-attachment	0.0118
22	veil-type	0
	$NT = 100,\ F = 3$	
1	odor	0.518
2	bruises	0.391
3	stalk-root	0.364
4	habitat	0.36
5	gill-size	0.338
6	spore-print-color	0.334
7	gill-spacing	0.334
8	stalk-shape	0.327
9	cap-color	0.313
10	cap-surface	0.306
11	population	0.297
12	ring-type	0.291
13	gill-color	0.265
14	cap-shape	0.249
15	stalk-color-below-ring	0.246
16	ring-number	0.241
17	stalk-surface-below-ring	0.212
18	stalk-surface-above-ring	0.209
19	veil-color	0.177
20	stalk-color-above-ring	0.162
21	gill-attachment	0.0616
22	veil-type	0
	$NT = 100,\ F = int(log_2(M))$	
1	odor	0.501
2	stalk-root	0.423
3	bruises	0.402
4	ring-type	0.363
5	cap-surface	0.36
6	spore-print-color	0.356

7	gill-spacing	0.35
8	gill-size	0.319
9	population	0.319
10	habitat	0.314
11	gill-color	0.303
12	stalk-shape	0.299
13	stalk-surface-below-ring	0.292
14	stalk-surface-above-ring	0.28
15	cap-color	0.278
16	cap-shape	0.257
17	ring-number	0.242
18	stalk-color-below-ring	0.218
19	veil-color	0.212
20	stalk-color-above-ring	0.185
21	gill-attachment	0.0785
22	veil-type	0
	$NT = 100,\ F = \sqrt{M}$	
1	odor	0.482
2	stalk-root	0.388
3	ring-type	0.368
4	habitat	0.355
5	bruises	0.354
6	gill-size	0.35
7	population	0.334
8	cap-surface	0.324
9	gill-spacing	0.316
10	spore-print-color	0.315
11	cap-shape	0.309
12	cap-color	0.307
13	stalk-surface-below-ring	0.291
14	stalk-shape	0.288
15	gill-color	0.286
16	stalk-surface-above-ring	0.278
17	ring-number	0.257

18	stalk-color-below-ring	0.22
19	stalk-color-above-ring	0.211
20	veil-color	0.206
21	gill-attachment	0.115
22	veil-type	0

Car

	$NT = 50,\ F = 1$	
1	safety	0.494
2	buying	0.484
3	maint	0.463
4	persons	0.404
5	doors	0.391
6	lug_boot	0.376
	$NT = 50,\ F = 3$	
1	doors	0.548
2	maint	0.532
3	lug_boot	0.517
4	buying	0.508
5	persons	0.498
6	safety	0.487
	$NT = 50,\ F = int(log_2(M))$	
1	buying	0.524
2	safety	0.505
3	maint	0.505
4	lug_boot	0.468
5	doors	0.459
6	persons	0.448
	$NT = 50,\ F = \sqrt{M}$	
1	safety	0.515
2	maint	0.509
3	buying	0.509
4	lug_boot	0.509

5	doors	0.458
6	persons	0.445
	$NT = 100,\ F = 1$	
1	safety	0.511
2	buying	0.477
3	maint	0.468
4	persons	0.379
5	doors	0.378
6	lug_boot	0.375
	$NT = 100,\ F = 3$	
1	doors	0.543
2	lug_boot	0.526
3	maint	0.518
4	buying	0.514
5	persons	0.477
6	safety	0.464
	$NT = 100,\ F = int(log_2(M))$	
1	maint	0.512
2	lug_boot	0.502
3	safety	0.499
4	buying	0.496
5	doors	0.468
6	persons	0.46
	$NT = 100,\ F = \sqrt{M}$	
1	maint	0.515
2	safety	0.509
3	buying	0.5
4	lug_boot	0.489
5	persons	0.487
6	doors	0.465