EE542 Lab 10
GDC Data

YONG WANG, YONGW@USC.EDU

YUHAO ZHANG, YUHAOZ@USC.EDU

Part 1: Data processing

1. Download files of all the disease types from GDC
2. Check integrity
3. Generate relationship between file_id and case_id
4. Retrieve file and case meta data from GDC repository
5. Generate relationship between miRNA and primary site (label)

Download miRNA files from GDC

Select all disease type in ‘Cases’

Select miRNA-Seq in ‘Files’

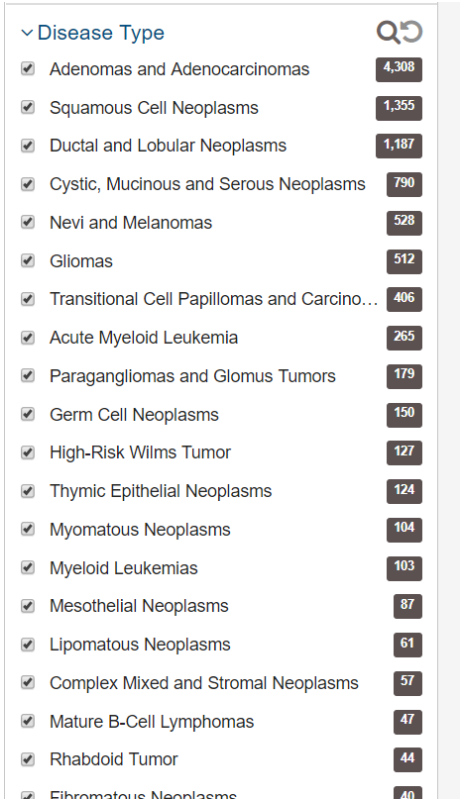
Download Manifest file

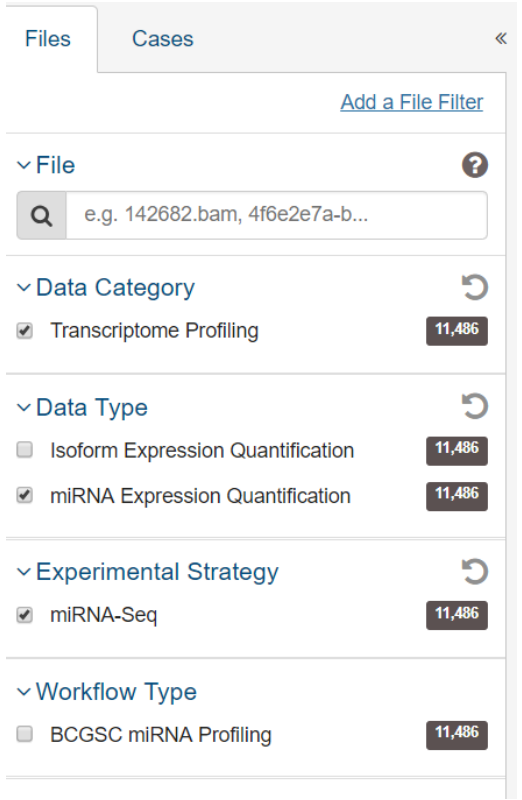
Download JSON file

Use gdc-client and manifest file to download all the files

Total 11486 files

```
100% [##################################################]
Successfully downloaded: 11486
wy@instance-2:~/ee542lab10$ ls
```





Check Integrity

Run check.py to check all downloaded files

```
wy@instance-2:~/ee542lab10/src$ python3 check.py
[2018-10-18 20:59:06,417 - GDC - INFO] ====start checking====
[2018-10-18 20:59:24,562 - GDC - INFO] successful downloads
[2018-10-18 20:59:24,563 - GDC - INFO] ====check finished====
wy@instance-2:~/ee542lab10/src$
```

Generate Relationship file_id and case_id

Run parse_file_case_id.py

This program use JSON file as input to generate relationship between file_id and case_id.

	A	B	C
1	Column1	Column2	
2	file_id	case_id	
3	52792628-b48a-4068-878c-960d5f6ebbcd	5a2f8140-8f90-4e94-b703-5fa5aa96be7b	
4	eb73f062-1fe7-484e-ab22-856d2fd936ad	bebd0025-74e2-451b-93b3-86f82df43573	
5	a51d1524-3daa-4f38-b9ea-d705bf96f08e	82476d2d-e403-4f6b-8dd6-cc84e3329478	
6	143f51d7-474a-4b4d-872c-7ccf6fcf1eb3	63bd2175-4b7c-44c9-aef3-9efc8f79837b	
7	824744e2-be27-4ea9-994e-8e53083033d0	c0fdb152-25d2-404b-bec7-a43ece381f5b	

Retrieve file and case meta data

Run request_meta.py

This program generate files_meta.tsv, cases_meta.tsv

	A	B	C	
1	cases.0.samples.0.portions.0.analytes.0.aliquots.0.aliquot_id	data_type	cases.0.samples.0.sample_type	file_
2	4aaff894-0a42-4b20-9290-759a34e6f248	miRNA Expression Quantification	Primary Tumor	9a2
3	f4a61521-e301-4b9a-8958-e483136011d1	miRNA Expression Quantification	Primary Tumor	955
4	874c3ea2-4ea4-467d-8c31-49ec735bd4f2	miRNA Expression Quantification	Primary Tumor	805
5	ad9ebe65-3dce-4b52-909c-99b07343d814	miRNA Expression Quantification	Primary Tumor	7ae
6	4eededac-03b6-4072-b518-b6760da4e656	miRNA Expression Quantification	Primary Tumor	37b
7	e15c3854-997c-4197-877f-99ac7695cb3a	miRNA Expression Quantification	Primary Tumor	69c

DD	DE	DF	DG	
demographic.updated_datetime	primary_site	exposures.0.years_smoked	submitter_aliquot_ids.46	s
9/6/2018 13:49	Breast			
9/6/2018 13:49	Breast			
9/6/2018 13:49	Breast			
9/6/2018 13:49	Breast			
9/6/2018 13:49	Breast			
9/6/2018 13:49	Breast			

Generate relationship between miRNA and primary site(label)

Run gen_miRNA_matrix.py to generate file_case_id_miRNA.csv

```
wy@instance-2:~/ee542lab10/src$ python3 gen_miRNA_matrix.py
[2018-10-20 01:31:44,050 - GDC - INFO] 691 Normal samples, 10790 Tumor samples
```

BTI	BTJ	BTK
hsa-mir-99a	hsa-mir-99b	label
249	47138	14
5958	201866	22
996	77803	35
1241	42229	29
8	55894	11
1576	34120	1
8012	215998	31
1987	360036	13
697	59606	16
2075	58314	0
4519	80501	1

Left plot shows miRNA and label

	A	B
1	primary_site	label
2	Normal	0
3	Breast	1
4	Bronchus and lung	2
5	Larynx	3
6	Retroperitoneum and peritoneum	4
7	Uterus, NOS	5
8	Connective, subcutaneous and other soft tis	6
9	Kidney	7
10	Cervix uteri	8

Right plot shows label and primary site

Part2 Apply sklearn to Data

1. Data Standardization
2. Split into training and test data (70%, 30%)
3. Feature selection
4. Turning hyper-parameters with Cross validation
5. Evaluation

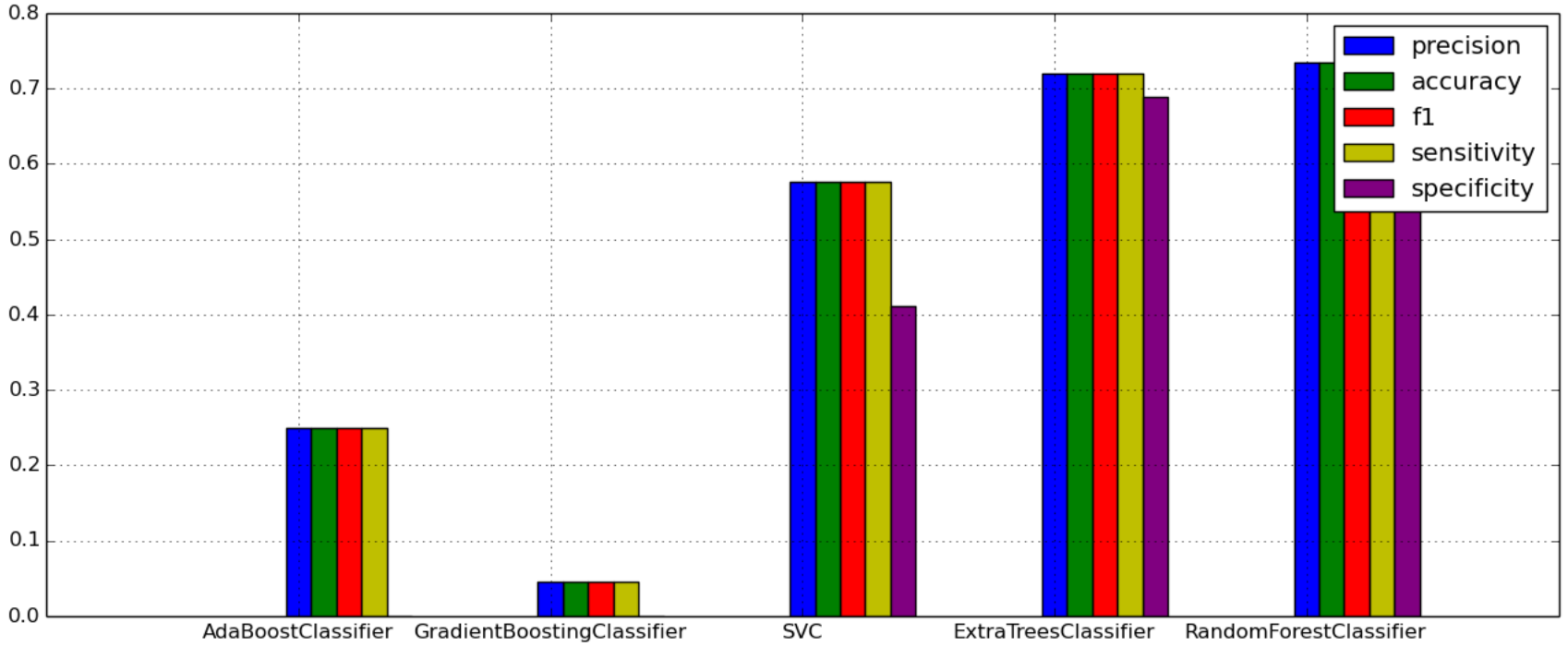
Feature selection

1. Use SelectFromModel and LassoCV to select 20 features
2. Compare different models
3. Select the best model and increase selected features

```
[2018-10-20 05:05:10,798 - GDC - INFO] selected features are [10, 26, 78, 119, 195, 240, 306,
352, 497, 498, 515, 539, 588, 991, 1148, 1461, 1665, 1722, 1848, 1872]
```

Compare different models

Use 20 selected features



Tuning RandomForest Model

Increase number of selected feather to 190

Tuning n_estimater to [100, 500] for RandomForestClassifier

Result:

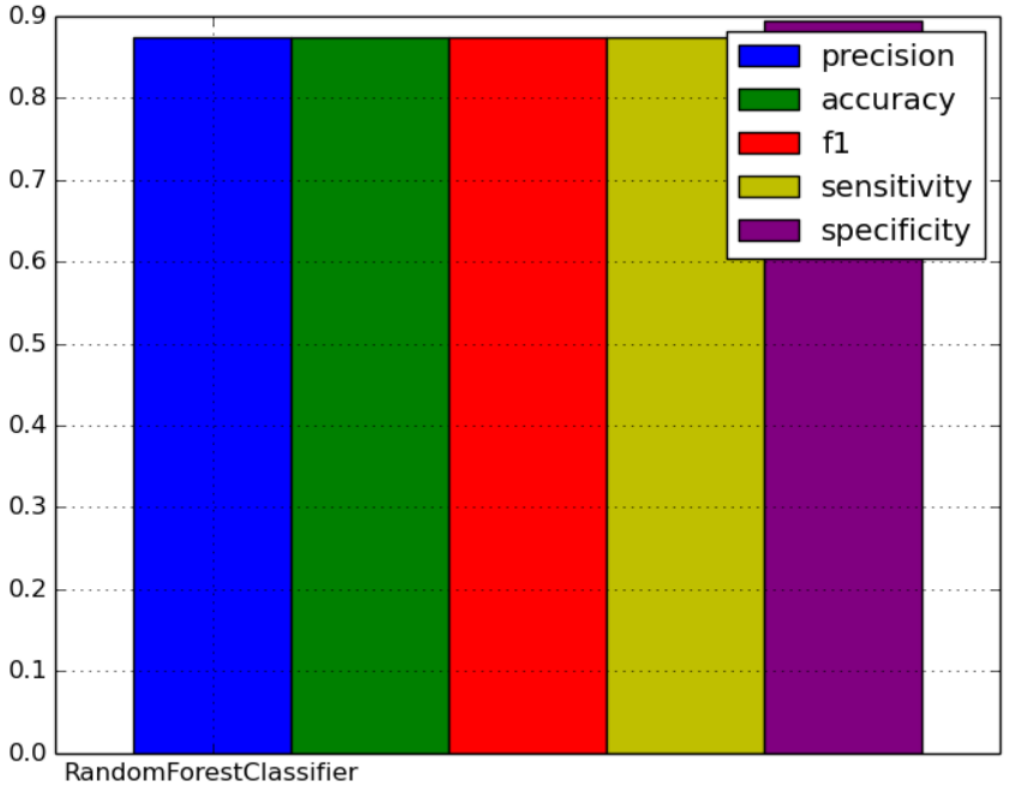
```
[2018-10-20 08:44:32,622 - GDC - INFO] Percentage of tumor cases in training set is 0.9400199104031857

[2018-10-20 08:44:32,622 - GDC - INFO] Percentage of tumor cases in test set is 0.93933236574746

[2018-10-20 08:57:28,619 - GDC - INFO] selected features are [4, 5, 8, 10, 13, 26, 49, 72, 78, 88, 90, 92,
93, 119, 141, 175, 180, 191, 194, 195, 201, 203, 204, 229, 232, 233, 239, 240, 245, 248, 249, 255, 264, 266,
270, 272, 273, 286, 296, 299, 302, 304, 305, 306, 309, 325, 327, 329, 332, 339, 344, 352, 381, 387, 406,
426, 429, 448, 458, 464, 470, 477, 482, 483, 492, 493, 495, 496, 497, 498, 500, 503, 505, 513, 514, 515,
532, 539, 544, 566, 588, 593, 595, 615, 623, 633, 638, 645, 646, 676, 677, 680, 692, 710, 764, 777, 784,
810, 813, 814, 834, 836, 847, 860, 880, 884, 888, 894, 900, 907, 911, 950, 957, 958, 969, 991, 996, 1004,
1038, 1041, 1048, 1063, 1072, 1078, 1079, 1091, 1102, 1111, 1135, 1141, 1148, 1152, 1232, 1251, 1267, 1274,
1289, 1305, 1316, 1337, 1342, 1362, 1363, 1364, 1369, 1376, 1378, 1387, 1395, 1402, 1406, 1410, 1412, 1447,
1461, 1475, 1487, 1504, 1507, 1509, 1516, 1524, 1544, 1546, 1560, 1584, 1588, 1638, 1644, 1665, 1677, 1695,
1717, 1720, 1722, 1733, 1747, 1750, 1771, 1786, 1791, 1834, 1843, 1848, 1859, 1860, 1872, 1874, 1875, 1879]

[2018-10-20 19:05:41,541 - GDC - INFO] scores are {'RandomForestClassifier': [0.8737300435413643,
0.8737300435413643, 0.87373004354136441, 0.8737300435413643, 0.894736842105263 2]}

{'n_estimators': 500}
```

Evaluation

190 selected features (total: 1881 features)
11481 records (70% train, 30% test)
55 classes

Model: RandomForestClassifier
Hyper-parameter: n_estimaters: 500

Precision: 0.87373
Accuracy: 0.87373
F1: 0.87373
Recall: 0.87373
Specificity: 0.89474



Code

https://github.com/uscwy/ee542lab10

Reference

The Pandas DataFrame – loading, editing, and viewing data in Python

Pandas manual

Numpy manual

Sklearn manual

Thanks