TECHNISCHE UNIVERSITÄT MÜNCHEN

Guided Research in Bioinformatics

Data Processing, Data Mining and Machine Learning in the field of Cancer Genomics

Sohel Mahmud

Outlines

- Motivation
- Goals
- Data Inspection
- Descriptive Mining
- Predictive Mining

Motivation

- The Cancer Genome Atlas (TCGA) is a project under the supervision of the National Cancer Institute's Center for Cancer Genomics and the National Human Genome Research Institute
- Seven Genome data analysis centers(GDAC) funded by the NCI/NHGRI are responsible for the integration of data across all characterization and sequencing centers as well as biological interpretation of TCGA data.
- All seven GDACs work together to develop an analysis pipeline for automated data analysis [3].
- TCGA has already 38 different types of cancer data including 10 rare cancer genomic data [1].

Goals

- to make access and understand The Cancer Genome Atlas data easier.
- to review the current data formats and interfaces for the access.
- to analyze the cancer patient data and summarize it using statistical analysis.
- to predict the cancer types with the TCGA data applying machine learning algorithms.

Data Inspection

Data Overview

- In this experiment, the data set has been acquired from Broad Institute of MIT and Harvard, a Genome Data Analysis Center (GDAC)
- Data set url: http://gdac.broadinstitute.org/runs/stddata__2016_01_28/data/
- contains 36 types of tumors.
- Throughout the data analysis process, 4 fields have been considered:
  - Tumor name: collected from the directory name
  - Patient ID: collected from the name of .maf files
  - HUGO Symbol: collected from the .maf files
  - Variant Classification: collected from the .maf files

Data Overview

- 6 different types of tumors were not available during the collection of data set,
  - Clear Cell Sarcoma of the Kidney (CCSK)
  - High-Risk Wilms Tumor (WT)
  - Neuroblastoma (NBL)
  - Osteosarcoma (OS)
  - Rhabdoid Tumor (RT)
  - Mesothelioma (MESO)
- Moreover, 9 tumors have been clustered into 4 different tumor super classes:
  - Glioma (GBMLGG = GBM + LGG)
  - Colorectal adenocarcinoma (COADREAD = COAD + READ)
  - Pan-kidney cohort (KIPAN = KICH + KIRC + KIRP)
  - Stomach and Esophageal carcinoma (STES = ESCA + STAD).
- However, these 9 tumors still exist in the data set, which induces the duplication of the patients' data.

Duplication of Patients data

Tumor Name	# of Patients	New tumor Class	# of Patients
Colon adenocarcinoma (COAD)	154	Colorectal adenocarcinoma (COADREAD)	223
Rectum adenocarcinoma (READ)	69		
Glioblastoma multiforme (GBM)	290	Glioma (GBMLGG)	576
Brain Lower Grade Glioma (LGG)	286		
Esophageal carcinoma (ESCA)	185	Stomach and Esophageal carcinoma (STES)	474
Stomach adenocarcinoma (STAD)	289		
Kidney Chromophobe (KICH)	66	Pan-kidney cohort (KIPAN)	644
Kidney renal clear cell carcinoma (KIRC)	417		
Kidney renal papillary cell carcinoma (KIRP)	161		

Summary of Genomic Data Set

- The summary of the cancer data set before removing the duplicate entries:

# of Tumors	# of Patients	Total Mutated Genes(overlapped)	Total Mutated Genes(unique)	Total Mutations	# of Mutation types
36	9037	352,570	22,255	1,357,856	22

- After removing the similar patients from the data set, the final figures are as follows:

# of Tumors	# of Patients	Total Mutated Genes(overlapped)	Total Mutated Genes(unique)	Total Mutations	# of Mutation types
27	7,120	263,336	22,255	1,068,351	22

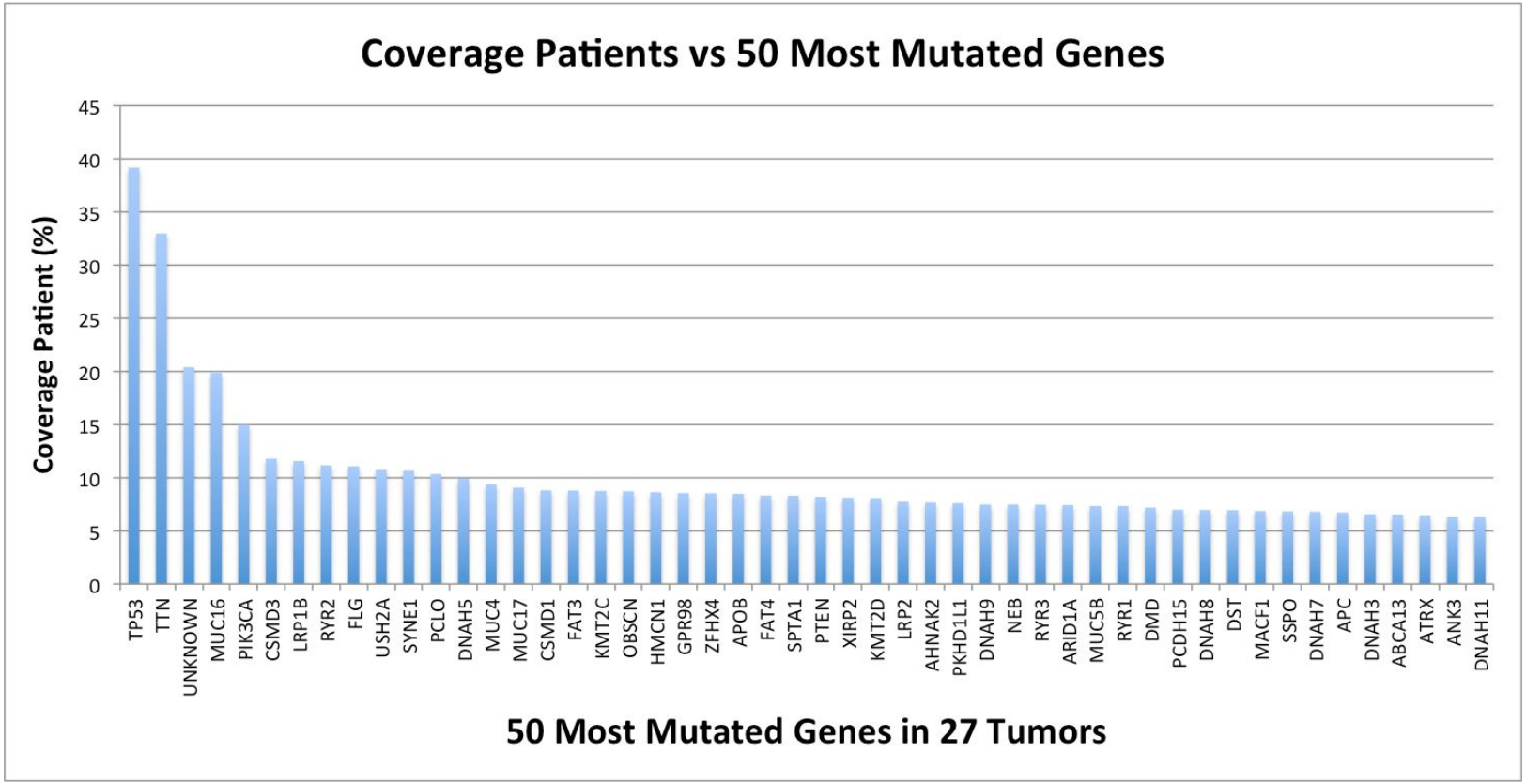
Descriptive Mining

Top 5 Frequent Mutation Types

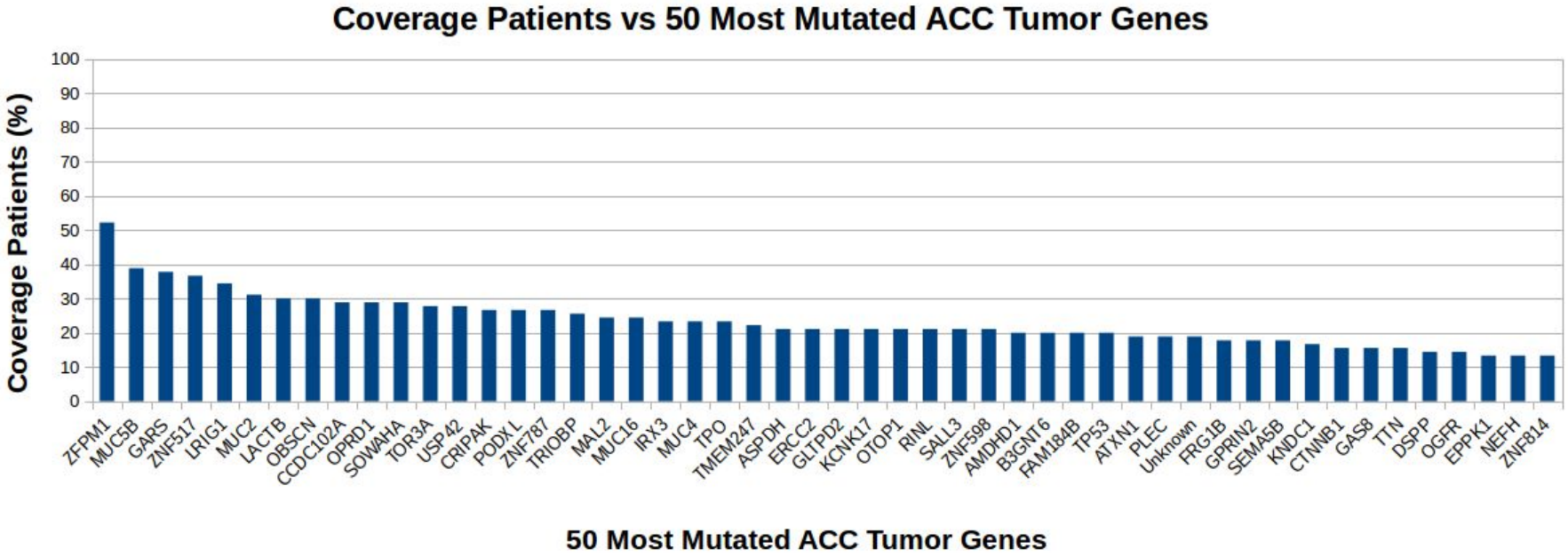
Top 5 Frequent Mutation Types with Frequencies:

Type of Mutations	Frequency	Percentage (%)
Missense_Mutation	1049456	77.288
Nonsense_Mutation	78206	5.760
Splice_Site	52427	3.861
Frame_Shift_Del	49188	3.622
Intron	27322	2.012

Patient Coverage of 50 Most Mutated Genes



50 Most Mutated Genes in ACC Tumor



Outcomes

- The large tumor sample is **BRCA** (Breast invasive carcinoma) consists of 982 patients.
- The number of highest mutated gene is found in **STES** (Stomach and Esophageal carcinoma) tumor having **474** patients with **18821** mutated genes.
- Tumor **SKCM** (Skin Cutaneous Melanoma) has highest mutations 198,846 in 17660 mutated genes for **345** patients and **11** mutations per affected gene.
- **DLBC** (Lymphoid Neoplasm Diffuse Large B-cell Lymphoma) has highest number of mutated genes per patient (**136** mutated genes per patient).
- Among 50 most frequently mutated genes for 27 types of tumor, the most frequently mutated gene is **TP53** (mutated in 33% of all patients).
- The second highly mutated gene is **TTN**, and it is found almost 28% of the patients.
- The top most mutated gene in ACC tumor is ZFPM1 (mutated 52.2% in ACC patients).
- The ratio of the mutation frequency of ZFPM1 to the total number of mutations is 112 : 13955. This means in 13955 mutations, ZFPM1 is mutated only 112 times which is 0.8%.
- In the tumor ACC, the total number of mutations is 13955, and the total number of mutated genes is 7108, and the ratio is 1: 1.96 ≈ 1: 2.

Predictive Mining

Applications of Machine Learning Algorithms

Feature Selection

- 50 most frequently mutated genes have been chosen from 27 tumors.
- the overlapped genes are filtered to 613 from 1,800 genes (50 genes from each tumor)
- A binary matrix has been formed with 614 (613 genes and tumor class) attributes for 7,120 cancer patients.
- The instances having no mutated gene are removed from the matrix.
- Finally, the binary matrix has 614 attributes and 6998 instances.
- For predicting the tumor types, the machine-learning algorithms are:
  - Naive Bayes Classifier
  - Decision Trees (J48)
  - Rule-based classification (PART)
- Test mode: 10-fold cross-validation
- All prediction tasks have been accomplished by using Waikato Environment for Knowledge Analysis (WEKA version-3.8.0).

Prediction Results of Machine Learning Algorithms

Summary of all classifiers prediction result:

Classifiers	Correctly Classified	Incorrectly Classified
Naive Bayes	3529 (50.4287 %)	3469 (49.5713%)
Decision Trees (J48)	3141 (44.8843 %)	3857 (55.1157 %)
Rule-Based (PART)	3187 (45.5416 %)	3811 (54.4584 %)

Naive Bayes Classifier

Evaluation of Naïve Bayes Classifier:

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC-Area	Class	Status
0.925	0.003	0.779	0.925	0.846	0.985	UVM	Highest
0.895	0.005	0.614	0.895	0.729	0.946	UCS	2nd Highest
0.276	0.025	0.282	0.276	0.279	0.841	SARC	Lowest
0.504	0.026	0.550	0.504	0.518	0.874		Wgt Average

Naive Bayes Classifier

Confusion Matrix

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z	aa	<-- classified as
71	1	1	2	0	0	1	0	0	3	0	1	0	0	0	0	0	0	1	3	2	0	0	0	4	0	0	a = ACC(90)
0	41	2	7	0	5	0	0	10	2	0	6	0	3	0	0	0	0	0	4	25	0	0	0	25	0	0	b = BLCA(130)
0	12	474	73	0	23	1	29	36	35	5	61	2	0	61	3	8	29	33	7	25	8	1	1	31	2	4	c = BRCA(982)
0	2	12	77	0	6	0	0	6	6	0	12	1	0	0	3	1	4	7	5	19	2	0	3	22	2	0	d = CESC(194)
0	0	0	7	16	0	1	0	1	0	0	1	1	0	0	0	0	0	0	0	5	0	0	0	3	0	0	e = CHOL(35)
0	0	9	1	0	132	0	0	3	0	0	6	0	0	9	22	0	0	4	16	10	0	0	0	11	0	0	f = COADREAD(223)
0	1	2	2	0	0	23	0	2	0	0	2	0	0	0	0	0	1	2	1	3	1	0	0	8	0	0	g = DLBC(48)
0	8	19	59	4	18	3	342	25	10	4	22	3	1	4	0	9	9	10	0	5	2	3	1	4	4	0	h = GBMLGG(576)
0	16	14	14	0	16	0	4	109	1	0	8	6	13	8	2	0	2	5	9	40	0	0	0	11	1	0	i = HNSC(279)
11	16	34	96	3	11	4	9	14	329	3	35	0	0	1	5	4	15	19	0	2	17	0	2	6	4	2	j = KIPAN(644)
0	0	6	0	0	0	0	3	1	12	132	1	0	0	10	1	3	9	3	0	0	5	2	0	0	0	0	k = LAML(197)
0	5	8	14	0	18	0	5	11	4	1	75	3	0	1	4	0	2	6	2	20	0	0	0	17	0	1	l = LIHC(198)
0	1	1	12	0	15	1	9	13	5	0	7	79	7	2	5	3	3	1	41	18	1	1	0	3	2	0	m = LUAD(230)
0	3	0	1	0	5	0	1	9	1	0	1	11	58	1	1	1	0	0	17	59	0	0	0	8	1	0	n = LUSC(178)
0	6	29	2	0	22	0	13	41	6	0	35	2	0	121	2	0	7	16	0	7	2	1	2	0	2	0	o = OV(316)
0	1	3	2	0	25	0	0	2	0	0	5	0	0	1	108	0	0	0	1	2	0	0	0	0	0	0	p = PAAD(150)
0	1	16	0	0	0	0	3	0	8	1	3	0	0	0	0	115	14	8	0	0	8	1	3	0	0	0	q = PCPG(184)
0	4	72	9	0	1	1	13	1	14	0	23	0	0	14	0	8	101	27	1	1	3	4	2	2	0	1	r = PRAD(332)
0	2	25	23	0	2	0	13	14	6	3	26	2	0	10	4	5	15	67	4	3	4	0	10	1	1	3	s = SARC(247)
0	1	1	10	1	13	0	4	8	8	1	2	30	1	1	0	4	0	0	222	13	13	7	0	1	0	4	t = SKCM(345)
0	8	21	21	1	41	0	2	41	2	0	21	7	11	4	6	0	2	9	36	190	0	0	0	46	3	1	u = STES(474)
0	1	1	22	2	0	0	1	0	11	0	6	0	0	0	0	3	1	0	1	0	100	0	0	0	6	0	v = TGCT(155)
0	0	17	0	0	0	0	1	0	6	5	5	0	0	2	4	13	37	10	0	0	0	270	1	0	0	4	w = THCA(405)
0	0	5	1	0	0	0	2	0	3	2	1	0	0	1	1	3	19	7	1	0	0	0	65	0	0	1	x = THYM(123)
0	0	22	40	0	1	0	28	11	2	0	25	0	0	3	1	0	1	2	15	3	1	0	0	87	4	0	y = UCEC(248)
0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1	0	0	2	1	0	0	0	0	51	0	z = UCS(57)
0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	2	1	0	0	0	0	0	1	0	74	aa = UVM(80)

Naive Bayes Classifier

Confusion Matrix

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z	aa	<-- classified as
71	1	1	2	0	0	1	0	0	3	0	1	0	0	0	0	0	0	1	3	2	0	0	0	4	0	0	a = ACC(90)
0	41	2	7	0	5	0	0	10	2	0	6	0	3	0	0	0	0	0	4	25	0	0	0	25	0	0	b = BLCA(130)
0	12	474	73	0	23	1	29	36	35	5	61	2	0	61	3	8	29	33	7	25	8	1	1	31	2	4	c = BRCA(982)
0	2	12	77	0	6	0	0	6	6	0	12	1	0	0	3	1	4	7	5	19	2	0	3	22	2	0	d = CESC(194)
0	0	0	7	16	0	1	0	1	0	0	1	1	0	0	0	0	0	0	0	5	0	0	0	3	0	0	e = CHOL(35)
0	0	9	1	0	132	0	0	3	0	0	6	0	0	9	22	0	0	4	16	10	0	0	0	11	0	0	f = COADREAD(223)
0	1	2	2	0	0	23	0	2	0	0	2	0	0	0	0	0	1	2	1	3	1	0	0	8	0	0	g = DLBC(48)
0	8	19	59	4	18	3	342	25	10	4	22	3	1	4	0	9	9	10	0	5	2	3	1	4	4	0	h = GBMLGG(576)
0	16	14	14	0	16	0	4	109	1	0	8	6	13	8	2	0	2	5	9	40	0	0	0	11	1	0	i = HNSC(279)
11	16	34	96	3	11	4	9	14	329	3	35	0	0	1	5	4	15	19	0	2	17	0	2	6	4	2	j = KIPAN(644)
0	0	6	0	0	0	0	3	1	12	132	1	0	0	10	1	3	9	3	0	0	5	2	0	0	0	0	k = LAML(197)
0	5	8	14	0	18	0	5	11	4	1	75	3	0	1	4	0	2	6	2	20	0	0	0	17	0	1	l = LIHC(198)
0	1	1	12	0	15	1	9	13	5	0	7	79	7	2	5	3	3	1	41	18	1	1	0	3	2	0	m = LUAD(230)
0	3	0	1	0	5	0	1	9	1	0	1	11	58	1	1	1	0	0	17	59	0	0	0	8	1	0	n = LUSC(178)
0	6	29	2	0	22	0	13	41	6	0	35	2	0	121	2	0	7	16	0	7	2	1	2	0	2	0	o = OV(316)
0	1	3	2	0	25	0	0	2	0	0	5	0	0	1	108	0	0	0	1	2	0	0	0	0	0	0	p = PAAD(150)
0	1	16	0	0	0	0	3	0	8	1	3	0	0	0	0	115	14	8	0	0	8	1	3	0	0	0	q = PCPG(184)
0	4	72	9	0	1	1	13	1	14	0	23	0	0	14	0	8	101	27	1	1	3	4	2	2	0	1	r = PRAD(332)
0	2	25	23	0	2	0	13	14	6	3	26	2	0	10	4	5	15	67	4	3	4	0	10	1	1	3	s = SARC(247)
0	1	1	10	1	13	0	4	8	8	1	2	30	1	1	0	4	0	0	222	13	13	7	0	1	0	4	t = SKCM(345)
0	8	21	21	1	41	0	2	41	2	0	21	7	11	4	6	0	2	9	36	190	0	0	0	46	3	1	u = STES(474)
0	1	1	22	2	0	0	1	0	11	0	6	0	0	0	0	3	1	0	1	0	100	0	0	0	6	0	v = TGCT(155)
0	0	17	0	0	0	0	1	0	6	5	5	0	0	2	4	13	37	10	0	0	0	270	1	0	0	4	w = THCA(405)
0	0	5	1	0	0	0	2	0	3	2	1	0	0	1	1	3	19	7	1	0	0	0	65	0	0	1	x = THYM(123)
0	0	22	40	0	1	0	28	11	2	0	25	0	0	3	1	0	1	2	15	3	1	0	0	87	4	0	y = UCEC(248)
0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1	0	0	2	1	0	0	0	0	51	0	z = UCS(57)
0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	2	1	0	0	0	0	0	1	0	74	aa = UVM(80)

Decision Trees (J48)

- Evaluation of Decision Trees (J48) Classifier:

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC-Area	Class	Status
0.875	0.003	0.769	0.875	0.819	0.939	UVM	Highest
0.717	0.015	0.733	0.717	0.725	0.892	THCA	2nd Highest
0.029	0.003	0.043	0.029	0.034	0.527	CHOL	Lowest
0.449	0.034	0.445	0.449	0.444	0.758		Wgt Average

Decision Trees (J48)

Confusion Matrix

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z	aa	<-- classified as
59	2	5	2	0	1	1	0	3	7	1	3	0	1	0	0	0	2	0	1	0	2	0	0	0	0	0	a = ACC(90)
3	18	13	6	0	2	0	7	14	14	0	11	3	8	7	0	0	4	0	3	10	1	4	1	1	0	0	b = BLCA(130)
5	17	442	16	0	8	6	40	29	45	8	22	7	14	95	4	19	82	28	4	45	5	9	4	9	0	1	c = BRCA(982)
2	6	28	29	0	4	2	7	8	22	8	1	8	2	0	4	12	5	7	4	8	8	1	6	5	0	3	d = CESC(194)
0	1	0	2	1	1	0	2	0	3	5	1	2	2	0	0	2	0	1	5	2	5	0	0	0	0	0	e = CHOL(35)
3	3	5	2	0	130	0	2	6	4	0	8	4	5	4	8	1	1	1	8	15	0	6	0	6	0	1	f = COADREAD(223)
1	2	4	0	0	2	14	2	5	2	0	4	1	0	0	1	1	4	0	1	1	0	0	1	1	0	1	g = DLBC(48)
2	5	45	5	0	5	2	369	5	19	10	5	8	6	8	0	15	17	15	0	11	0	5	3	9	0	0	h = GBMLGG(576)
3	21	32	2	0	1	3	13	82	13	2	8	7	17	29	0	1	10	2	3	24	1	3	0	2	0	0	i = HNSC(279)
2	8	48	14	1	5	6	35	13	379	10	11	4	2	7	2	8	36	11	2	9	12	9	4	2	0	2	j = KIPAN(644)
0	1	8	1	2	0	1	13	0	9	104	1	1	0	11	4	1	9	3	7	0	7	4	0	1	0	0	k = LAML(197)
3	7	20	4	0	7	2	17	7	14	5	49	2	5	9	2	0	18	7	3	11	0	1	0	4	0	0	l = LIHC(198)
1	4	15	6	0	12	1	26	11	10	5	6	56	10	2	15	4	2	2	10	14	7	4	2	3	1	1	m = LUAD(230)
3	11	20	7	2	6	0	7	32	4	1	10	7	28	10	0	0	1	1	2	24	0	0	0	2	0	0	n = LUSC(178)
2	9	50	1	0	9	1	12	15	13	3	6	2	8	139	1	0	10	6	1	18	1	2	0	6	0	1	o = OV(316)
0	1	6	2	0	5	1	0	1	2	0	2	12	0	1	106	0	1	1	0	7	0	2	0	0	0	0	p = PAAD(150)
0	1	15	1	3	0	1	4	1	12	1	0	1	0	1	0	81	24	3	0	3	7	13	8	0	1	0	q = PCPG(184)
2	0	47	5	1	1	4	17	5	24	0	12	0	2	13	0	4	135	7	1	3	3	11	3	2	0	0	r = PRAD(332)
1	5	21	9	2	2	0	21	8	15	3	7	3	3	24	1	13	29	40	1	15	2	2	13	2	0	1	s = SARC(247)
1	9	10	3	0	7	3	8	3	15	13	1	11	0	1	0	4	1	1	219	7	6	11	0	5	1	5	t = SKCM(345)
6	17	58	17	6	18	5	11	48	21	3	22	16	22	30	6	5	11	22	7	89	5	4	4	13	5	2	u = STES(474)
1	2	4	6	3	2	2	5	2	38	11	1	4	1	0	5	17	0	2	5	2	33	3	2	1	1	2	v = TGCT(155)
1	4	14	2	0	4	0	2	0	9	2	2	7	1	2	3	0	29	4	15	2	1	269	0	2	0	0	w = THCA(405)
0	2	5	7	0	0	0	1	1	3	2	1	0	0	1	1	5	19	5	0	2	1	3	53	0	0	0	x = THYM(123)
4	5	24	5	1	6	2	7	2	3	2	8	4	4	5	2	0	2	3	5	20	2	1	3	125	0	1	y = UCEC(248)
1	2	7	2	1	0	0	0	0	2	0	0	2	0	0	2	4	1	3	0	4	1	0	0	3	22	0	z = UCS(57)
0	0	2	0	0	0	0	1	0	2	0	1	0	0	0	0	0	2	0	0	2	0	0	0	0	0	70	aa = UVM(80)

Decision Trees (J48)

Confusion Matrix

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z	aa	<-- classified as
59	2	5	2	0	1	1	0	3	7	1	3	0	1	0	0	0	2	0	1	0	2	0	0	0	0	0	a = ACC(90)
3	18	13	6	0	2	0	7	14	14	0	11	3	8	7	0	0	4	0	3	10	1	4	1	1	0	0	b = BLCA(130)
5	17	442	16	0	8	6	40	29	45	8	22	7	14	95	4	19	82	28	4	45	5	9	4	9	0	1	c = BRCA(982)
2	6	28	29	0	4	2	7	8	22	8	1	8	2	0	4	12	5	7	4	8	8	1	6	5	0	3	d = CESC(194)
0	1	0	2	1	1	0	2	0	3	5	1	2	2	0	0	2	0	1	5	2	5	0	0	0	0	0	e = CHOL(35)
3	3	5	2	0	130	0	2	6	4	0	8	4	5	4	8	1	1	1	8	15	0	6	0	6	0	1	f = COADREAD(223)
1	2	4	0	0	2	14	2	5	2	0	4	1	0	0	1	1	4	0	1	1	0	0	1	1	0	1	g = DLBC(48)
2	5	45	5	0	5	2	369	5	19	10	5	8	6	8	0	15	17	15	0	11	0	5	3	9	0	0	h = GBMLGG(576)
3	21	32	2	0	1	3	13	82	13	2	8	7	17	29	0	1	10	2	3	24	1	3	0	2	0	0	i = HNSC(279)
2	8	48	14	1	5	6	35	13	379	10	11	4	2	7	2	8	36	11	2	9	12	9	4	2	0	2	j = KIPAN(644)
0	1	8	1	2	0	1	13	0	9	104	1	1	0	11	4	1	9	3	7	0	7	4	0	1	0	0	k = LAML(197)
3	7	20	4	0	7	2	17	7	14	5	49	2	5	9	2	0	18	7	3	11	0	1	0	4	0	0	l = LIHC(198)
1	4	15	6	0	12	1	26	11	10	5	6	56	10	2	15	4	2	2	10	14	7	4	2	3	1	1	m = LUAD(230)
3	11	20	7	2	6	0	7	32	4	1	10	7	28	10	0	0	1	1	2	24	0	0	0	2	0	0	n = LUSC(178)
2	9	50	1	0	9	1	12	15	13	3	6	2	8	139	1	0	10	6	1	18	1	2	0	6	0	1	o = OV(316)
0	1	6	2	0	5	1	0	1	2	0	2	12	0	1	106	0	1	1	0	7	0	2	0	0	0	0	p = PAAD(150)
0	1	15	1	3	0	1	4	1	12	1	0	1	0	1	0	81	24	3	0	3	7	13	8	0	1	0	q = PCPG(184)
2	0	47	5	1	1	4	17	5	24	0	12	0	2	13	0	4	135	7	1	3	3	11	3	2	0	0	r = PRAD(332)
1	5	21	9	2	2	0	21	8	15	3	7	3	3	24	1	13	29	40	1	15	2	2	13	2	0	1	s = SARC(247)
1	9	10	3	0	7	3	8	3	15	13	1	11	0	1	0	4	1	1	219	7	6	11	0	5	1	5	t = SKCM(345)
6	17	58	17	6	18	5	11	48	21	3	22	16	22	30	6	5	11	22	7	89	5	4	4	13	5	2	u = STES(474)
1	2	4	6	3	2	2	5	2	38	11	1	4	1	0	5	17	0	2	5	2	33	3	2	1	1	2	v = TGCT(155)
1	4	14	2	0	4	0	2	0	9	2	2	7	1	2	3	0	29	4	15	2	1	269	0	2	0	0	w = THCA(405)
0	2	5	7	0	0	0	1	1	3	2	1	0	0	1	1	5	19	5	0	2	1	3	53	0	0	0	x = THYM(123)
4	5	24	5	1	6	2	7	2	3	2	8	4	4	5	2	0	2	3	5	20	2	1	3	125	0	1	y = UCEC(248)
1	2	7	2	1	0	0	0	0	2	0	0	2	0	0	2	4	1	3	0	4	1	0	0	3	22	0	z = UCS(57)
0	0	2	0	0	0	0	1	0	2	0	1	0	0	0	0	0	2	0	0	2	0	0	0	0	0	70	aa = UVM(80)

Rule-based Classification (PART)

- Scheme: weka.classifiers.rules.PART -M 2 -C 0.25 -Q 1
- Evaluation of Rule-based (PART) Classifier:

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC-Area	Class	Status
0.9	0.002	0.818	0.9	0.857	0.948	UVM	Highest
0.715	0.015	0.73	0.715	0.722	0.9	THCA	2nd-Highest
0.114	0.002	0.2	0.114	0.145	0.608	CHOL	Lowest
0.455	0.035	0.448	0.455	0.45	0.76		Wgt Average

Rule-based Classification (PART)

Confusion Matrix

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z	aa	<-- classified as
59	5	7	1	0	1	2	2	1	4	0	5	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	a = ACC(90)
2	25	24	9	0	4	1	3	16	7	1	5	4	1	3	0	0	1	1	2	12	1	3	1	4	0	0	b = BLCA(130)
2	15	483	18	2	14	3	39	25	44	13	20	11	11	71	2	23	51	31	5	49	4	10	4	14	0	0	c = BRCA(982)
0	12	20	24	1	6	3	7	7	17	9	6	7	2	0	3	7	7	14	8	7	7	3	4	7	0	2	d = CESC(194)
0	1	0	4	4	0	0	2	0	4	1	0	2	0	0	0	3	0	2	3	0	8	0	0	1	0	0	e = CHOL(35)
3	5	4	2	0	131	0	5	7	3	2	7	7	5	3	5	0	2	1	8	13	0	2	0	7	0	1	f = COADREAD(223)
1	1	4	1	0	2	18	1	1	3	0	1	3	0	1	0	0	0	2	0	6	0	0	0	3	0	0	g = DLBC(48)
2	2	43	4	1	3	0	368	8	31	9	6	10	2	8	1	20	8	16	4	5	1	6	2	7	2	0	h = GBMLGG(576)
5	12	38	6	0	2	3	14	78	4	4	12	11	19	14	1	0	5	2	3	30	1	5	1	7	2	0	i = HNSC(279)
2	7	58	21	0	6	5	27	9	380	15	13	4	1	8	2	9	19	10	5	12	15	4	4	3	1	2	j = KIPAN(644)
1	0	8	0	0	0	0	14	3	3	113	0	1	1	8	4	2	9	2	4	2	7	6	0	0	0	0	k = LAML(197)
3	10	31	3	0	5	2	14	9	10	3	42	1	3	4	3	1	12	9	0	16	2	2	0	12	0	0	l = LIHC(198)
3	3	17	1	1	12	1	29	16	8	6	7	50	6	5	9	6	3	3	10	15	6	5	2	3	2	1	m = LUAD(230)
3	7	19	1	0	7	0	10	27	5	2	5	10	37	7	2	0	1	0	1	27	1	0	1	4	1	0	n = LUSC(178)
1	10	54	1	0	9	2	16	18	14	9	10	4	7	121	1	0	8	8	1	18	0	1	0	2	0	1	o = OV(316)
0	1	5	0	1	13	0	0	0	2	0	2	11	0	1	102	0	1	0	1	9	0	1	0	0	0	0	p = PAAD(150)
1	0	21	4	0	0	0	7	0	11	0	3	3	0	0	0	82	15	6	2	2	6	12	5	0	0	1	q = PCPG(184)
3	2	56	10	0	2	1	13	3	35	2	17	0	1	15	1	5	104	6	3	5	3	7	7	1	0	0	r = PRAD(332)
1	2	33	11	0	5	0	20	10	14	4	9	2	3	14	0	10	19	46	0	19	3	2	13	1	1	1	s = SARC(247)
1	3	6	1	2	8	3	8	4	11	12	2	9	4	4	1	4	3	1	206	12	10	20	0	5	1	4	t = SKCM(345)
10	15	56	12	4	26	3	9	33	19	2	19	15	19	22	6	4	11	21	8	121	5	2	0	26	3	2	u = STES(474)
3	0	7	4	1	2	1	3	3	31	17	0	1	1	0	3	18	2	3	4	2	41	2	2	2	1	1	v = TGCT(155)
1	2	13	0	0	3	0	4	1	12	6	2	1	1	2	4	1	24	10	19	0	0	268	0	1	0	0	w = THCA(405)
0	1	6	6	0	0	0	1	1	4	2	0	0	0	2	1	7	18	2	0	0	0	4	57	0	0	0	x = THYM(123)
0	3	15	1	3	6	3	10	4	10	4	4	6	1	8	2	0	3	6	3	13	6	1	3	131	0	0	y = UCEC(248)
0	1	3	0	0	2	0	2	1	1	0	0	0	1	0	2	5	0	3	0	6	1	0	1	4	24	0	z = UCS(57)
0	0	2	0	0	0	0	0	0	2	0	0	0	0	0	0	0	2	1	0	0	0	1	0	0	0	72	aa = UVM(80)

Rule-based Classification (PART)

Confusion Matrix

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z	aa	<-- classified as
59	5	7	1	0	1	2	2	1	4	0	5	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	a = ACC(90)
2	25	24	9	0	4	1	3	16	7	1	5	4	1	3	0	0	1	1	2	12	1	3	1	4	0	0	b = BLCA(130)
2	15	483	18	2	14	3	39	25	44	13	20	11	11	71	2	23	51	31	5	49	4	10	4	14	0	0	c = BRCA(982)
0	12	20	24	1	6	3	7	7	17	9	6	7	2	0	3	7	7	14	8	7	7	3	4	7	0	2	d = CESC(194)
0	1	0	4	4	0	0	2	0	4	1	0	2	0	0	0	3	0	2	3	0	8	0	0	1	0	0	e = CHOL(35)
3	5	4	2	0	131	0	5	7	3	2	7	7	5	3	5	0	2	1	8	13	0	2	0	7	0	1	f = COADREAD(223)
1	1	4	1	0	2	18	1	1	3	0	1	3	0	1	0	0	0	2	0	6	0	0	0	3	0	0	g = DLBC(48)
2	2	43	4	1	3	0	368	8	31	9	6	10	2	8	1	20	8	16	4	5	1	6	2	7	2	0	h = GBMLGG(576)
5	12	38	6	0	2	3	14	78	4	4	12	11	19	14	1	0	5	2	3	30	1	5	1	7	2	0	i = HNSC(279)
2	7	58	21	0	6	5	27	9	380	15	13	4	1	8	2	9	19	10	5	12	15	4	4	3	1	2	j = KIPAN(644)
1	0	8	0	0	0	0	14	3	3	113	0	1	1	8	4	2	9	2	4	2	7	6	0	0	0	0	k = LAML(197)
3	10	31	3	0	5	2	14	9	10	3	42	1	3	4	3	1	12	9	0	16	2	2	0	12	0	0	l = LIHC(198)
3	3	17	1	1	12	1	29	16	8	6	7	50	6	5	9	6	3	3	10	15	6	5	2	3	2	1	m = LUAD(230)
3	7	19	1	0	7	0	10	27	5	2	5	10	37	7	2	0	1	0	1	27	1	0	1	4	1	0	n = LUSC(178)
1	10	54	1	0	9	2	16	18	14	9	10	4	7	121	1	0	8	8	1	18	0	1	0	2	0	1	o = OV(316)
0	1	5	0	1	13	0	0	0	2	0	2	11	0	1	102	0	1	0	1	9	0	1	0	0	0	0	p = PAAD(150)
1	0	21	4	0	0	0	7	0	11	0	3	3	0	0	0	82	15	6	2	2	6	12	5	0	0	1	q = PCPG(184)
3	2	56	10	0	2	1	13	3	35	2	17	0	1	15	1	5	104	6	3	5	3	7	7	1	0	0	r = PRAD(332)
1	2	33	11	0	5	0	20	10	14	4	9	2	3	14	0	10	19	46	0	19	3	2	13	1	1	1	s = SARC(247)
1	3	6	1	2	8	3	8	4	11	12	2	9	4	4	1	4	3	1	206	12	10	20	0	5	1	4	t = SKCM(345)
10	15	56	12	4	26	3	9	33	19	2	19	15	19	22	6	4	11	21	8	121	5	2	0	26	3	2	u = STES(474)
3	0	7	4	1	2	1	3	3	31	17	0	1	1	0	3	18	2	3	4	2	41	2	2	2	1	1	v = TGCT(155)
1	2	13	0	0	3	0	4	1	12	6	2	1	1	2	4	1	24	10	19	0	0	268	0	1	0	0	w = THCA(405)
0	1	6	6	0	0	0	1	1	4	2	0	0	0	2	1	7	18	2	0	0	0	4	57	0	0	0	x = THYM(123)
0	3	15	1	3	6	3	10	4	10	4	4	6	1	8	2	0	3	6	3	13	6	1	3	131	0	0	y = UCEC(248)
0	1	3	0	0	2	0	2	1	1	0	0	0	1	0	2	5	0	3	0	6	1	0	1	4	24	0	z = UCS(57)
0	0	2	0	0	0	0	0	0	2	0	0	0	0	0	0	0	2	1	0	0	0	1	0	0	0	72	aa = UVM(80)

Conclusion

- Naïve Bayes classifier is very effective which accuracy rate was 50.4287 %
- Among 27 types of cancers, UVM (Uveal Melanoma) has been predicted efficiently
- SARC (Sarcoma), CHOL (Cholangiocarcinoma), and Cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC) have the lowest accurate prediction rate.

Reference

[1] Cancers Selected for Study - TCGA. (n.d.). Retrieved November 13, 2016, from https://cancergenome.nih.gov/cancersselected

[2] Rare Tumor Characterization Projects - TCGA. (n.d.). Retrieved November 13, 2016, from https://cancergenome.nih.gov/cancersselected/RareTumorCharacterizationProjects

[3] The Cancer Genome Atlas (2016, November 8). Retrived November 13, 2016, from https://en.wikipedia.org/wiki/The_Cancer_Genome_Atlas#Genome_data_analysis_centers

[4] MAF (Mutation Annotation Format). (n.d.). Retrieved November 13, 2016, from http://software.broadinstitute.org/software/igv/MutationAnnotationFormat

[5] Broad GDAC Firehose. (n.d.). Retrieved November 13, 2016, from https://gdac.broadinstitute.org/ [6] https://gdc.cancer.gov/about-data

Reference(contd.)

[6] Naive Bayes classifier. (2016, November 16). Retrieved November 13, 2016, from https://en.wikipedia.org/wiki/Naive_Bayes_classifier

[7] Witten IH, Frank E. Data mining: practical machine learning tools and techniques. Morgan Kaufmann; 2005

[8] F1-Measure. Retrieved November 13, 2016 from https://en.wikipedia.org/wiki/F1-measure

[9] Sayad, S. (n.d.). Naive Bayesian. Retrieved November 13, 2016, from http://www.saedsayad.com/naive_bayesian.htm

[10] Frank, E. (2013, March 15). How to see decision tree build for PART algorithm. Retrieved November 13, 2016, from http://weka.8497.n7.nabble.com/how-to-see-decison-tree-build-for-PART-algorithm-td27335.html

Acknowledgements

- Prof. Dr. Burkhard Rost for providing the opportunity to do the Guided Research as a supervisor.
- Dr. Lothar Richter for his patience, generosity and valuable suggestions throughout the whole research work.
- Yesmin Luna for her encouragement, enthusiasm and passionate cooperation throughout this whole journey.

Any Questions?

Thank you

Appendix

Genes Summary

Tumor Name	Number of Patients	Number of Mutated Genes	Total Number of Mutations	Mutations per Affected Gene	Mutations per patient	Mutated Genes per patient
ACC	90	7108	13955	2	155	79
BLCA	130	11774	29427	3	226	91
BRCA	982	16458	71437	4	72	17
CESC	194	14264	37004	3	190	74
CHOL	35	4098	5462	1	156	117
COADREAD	223	15043	65575	4	294	67
DLBC	48	6541	10918	2	227	136
GBMLGG	576	10245	24467	2	42	18
HNSC	279	12804	39263	3	140	46
KIPAN	644	13713	38447	3	59	21
LAML	197	1469	2113	1	10	7
LIHC	198	10239	21236	2	107	52
LUAD	230	14037	56809	4	246	61
LUSC	178	13807	49415	4	277	78
OV	316	8698	15971	2	50	28
PAAD	150	10193	22549	2	150	68
PCPG	184	2641	3758	1	20	14
PRAD	332	6281	9559	2	28	19
SARC	247	8839	16141	2	65	36
SKCM	345	17660	198846	11	1576	51
STES	474	18821	161016	9	339	40
TGCT	155	6851	11595	2	74	44
THCA	405	4008	5704	1	14	10
THYM	123	2098	2528	1	20	17
UCEC	248	18125	144170	8	581	73
UCS	57	6116	9297	2	163	107
UVM	80	1405	1689	1	21	18

Evaluation of Naive Bayes Classifier for 27 Tumors

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.925	0.003	0.779	0.925	0.846	0.985	UVM
	0.895	0.005	0.614	0.895	0.729	0.946	UCS
	0.789	0.002	0.866	0.789	0.826	0.938	ACC
	0.720	0.009	0.624	0.720	0.669	0.931	PAAD
	0.720	0.003	0.931	0.720	0.812	0.983	THCA
	0.702	0.004	0.841	0.702	0.765	0.984	LAML
	0.645	0.010	0.599	0.645	0.621	0.961	TGCT
	0.643	0.025	0.572	0.643	0.606	0.920	SKCM
	0.635	0.010	0.635	0.635	0.635	0.979	PCPG
	0.601	0.022	0.708	0.601	0.650	0.890	GBMLGG
	0.592	0.033	0.373	0.592	0.458	0.844	COADREAD
	0.580	0.004	0.722	0.580	0.644	0.965	THYM
	0.512	0.023	0.694	0.512	0.590	0.884	KIPAN
	0.492	0.053	0.596	0.492	0.539	0.836	BRCA
	0.479	0.002	0.657	0.479	0.554	0.868	DLBC
	0.457	0.002	0.593	0.457	0.516	0.956	CHOL
	0.405	0.061	0.156	0.405	0.225	0.765	CESC
	0.402	0.040	0.419	0.402	0.410	0.842	STES
	0.391	0.037	0.304	0.391	0.342	0.758	HNSC
	0.383	0.020	0.476	0.383	0.425	0.858	OV
	0.381	0.046	0.192	0.381	0.225	0.766	LIHC
	0.354	0.030	0.299	0.354	0.324	0.794	UCEC
	0.343	0.010	0.537	0.343	0.419	0.831	LUAD
	0.334	0.026	0.370	0.334	0.351	0.901	PRAD
	0.326	0.005	0.617	0.326	0.426	0.932	LUSC
	0.315	0.013	0.315	0.315	0.315	0.807	BLCA
	0.276	0.025	0.282	0.276	0.279	0.841	SARC
Weighted Avg.	0.504	0.026	0.550	0.504	0.518	0.874	

Evaluation of Decision Trees (J48) Classifier for 27 Tumors

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.875	0.003	0.769	0.875	0.819	0.939	UVM
	0.717	0.015	0.733	0.717	0.725	0.892	THCA
	0.707	0.009	0.635	0.707	0.669	0.862	PAAD
	0.656	0.007	0.557	0.656	0.602	0.827	ACC
	0.649	0.04	0.587	0.649	0.616	0.83	GBMLGG
	0.635	0.013	0.713	0.635	0.672	0.831	SKCM
	0.59	0.051	0.538	0.59	0.563	0.818	KIPAN
	0.583	0.016	0.546	0.583	0.564	0.816	COADREAD
	0.553	0.014	0.523	0.553	0.537	0.896	LAML
	0.508	0.012	0.613	0.508	0.556	0.782	UCEC
	0.473	0.008	0.495	0.473	0.484	0.85	THYM
	0.459	0.084	0.466	0.459	0.462	0.764	BRCA
	0.448	0.017	0.411	0.448	0.429	0.842	PCPG
	0.447	0.048	0.297	0.447	0.357	0.717	PRAD
	0.44	0.039	0.348	0.44	0.389	0.73	OV
	0.386	0.001	0.71	0.386	0.5	0.729	UCS
	0.294	0.033	0.272	0.294	0.283	0.684	HNSC
	0.292	0.006	0.246	0.292	0.267	0.656	DLBC
	0.249	0.023	0.241	0.249	0.245	0.659	LIHC
	0.243	0.017	0.326	0.243	0.279	0.666	LUAD
	0.213	0.011	0.3	0.213	0.249	0.73	TGCT
	0.188	0.04	0.256	0.188	0.217	0.605	STES
	0.165	0.02	0.229	0.165	0.191	0.688	SARC
	0.157	0.017	0.199	0.157	0.176	0.588	LUSC
	0.153	0.019	0.186	0.153	0.168	0.614	CESC
	0.138	0.021	0.11	0.138	0.123	0.585	BLCA
	0.029	0.003	0.043	0.029	0.034	0.527	CHOL
Weighted Avg.	0.449	0.034	0.445	0.449	0.444	0.758	

Evaluation of Rule-Based (PART) Classifier for 27 Tumors

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.9	0.002	0.818	0.9	0.857	0.948	UVM
	0.715	0.015	0.73	0.715	0.722	0.9	THCA
	0.68	0.008	0.658	0.68	0.669	0.869	PAAD
	0.656	0.007	0.551	0.656	0.599	0.846	ACC
	0.647	0.04	0.586	0.647	0.615	0.842	GBMLGG
	0.601	0.018	0.479	0.601	0.533	0.873	LAML
	0.597	0.014	0.687	0.597	0.639	0.833	SKCM
	0.592	0.049	0.552	0.592	0.571	0.818	KIPAN
	0.587	0.02	0.487	0.587	0.533	0.802	COADREAD
	0.533	0.017	0.535	0.533	0.534	0.796	UCEC
	0.509	0.007	0.533	0.509	0.521	0.844	THYM
	0.501	0.091	0.468	0.501	0.484	0.759	BRCA
	0.453	0.018	0.396	0.453	0.423	0.836	PCPG
	0.421	0.002	0.632	0.421	0.505	0.749	UCS
	0.383	0.03	0.377	0.383	0.38	0.737	OV
	0.375	0.005	0.353	0.375	0.364	0.693	DLBC
	0.344	0.033	0.317	0.344	0.33	0.718	PRAD
	0.28	0.031	0.274	0.28	0.277	0.647	HNSC
	0.265	0.013	0.313	0.265	0.287	0.721	TGCT
	0.256	0.043	0.302	0.256	0.277	0.636	STES
	0.217	0.018	0.289	0.217	0.248	0.657	LUAD
	0.213	0.023	0.213	0.213	0.213	0.64	LIHC
	0.208	0.013	0.294	0.208	0.243	0.629	LUSC
	0.192	0.017	0.172	0.192	0.182	0.609	BLCA
	0.189	0.024	0.223	0.189	0.205	0.672	SARC
	0.126	0.018	0.166	0.126	0.143	0.626	CESC
	0.114	0.002	0.2	0.114	0.145	0.608	CHOL
Weighted Avg.	0.455	0.035	0.448	0.455	0.45	0.76	