Prepoznavanje vrste kancera

Metoda klasifikacije

N. Bogdanović

Univerzitet u Beogradu:
Matematički Fakultet

13. septembar 2019

Plan izlaganja

Uputstvo za korišćenje

- Iz konzole pokrenuti main.py
- Priprema podataka za anlizu (procenjeno vreme: 2h)
- Analiziranje bez preprocesiranja:
  - Drveta odlučivanja: 5 - 10 min
  - Najbliži susedi: 20 - 30 min
  - Gausova metoda: 5 min
  - Neuronske mreže: 45 min

Podaci i alati

- https://portals.broadinstitute.org/ccle/data
- CCLE_ABSOLUTE_combined_20181227.xlsx
- ABSOLUTE_combined.segtab: 20 kolona i 188,653 redova
- segtab_annotations
- ABSOLUTE_combined.table
- data_original.xlsx
- Python: pandas, numpy, biblioteka IP-a

Atributi

- Sample
- Chromosome
- Start
- End
- Num_Probe
- Length
- Modal_HSCN_1
- Modal_HSCN_2
- Modal_HSCN_TOTAL
- Subclonal_HSCN_a1
- Subclonal_HSCN_a2
- Cancer_cell_frac_a1
- Ccf_ci95_low_a1
- Ccf_ci95_high_a1
- Cancer_cell_frac_a2
- Ccf_ci95_low_a2
- Ccf_ci95_high_a2
- LOH
- Homozygos_deletion
- depMapID

Problem prevelikog broja klasa

- LUNG: 38,882
- SALIVARY_GLAND: 358
- FIBROBLAST: 1,164
- PLEURA: 1,529
- THYROID: 2,824
- PANCREAS: 6,943
- BONE: 3,579
- INTESTINE: 9,369
- SOFT_TISSUE: 3,475
- ENDOMETRIUM: 5,277
- AUTNOMIC_GANGLIA: 2,115
- HAEMATOPOIETIC_AND_LYMPHOID_TISSUE: 27,115

Problem prevelikog broja klasa

- 5% tačnosti
- Najbrojnije klase: HAEMATOPOIETIC_AND_LYMPHOID_TISSUE i LUNG

Listing 1: Prečišćavanje klasa

```
i = 0
y = df["sample"]
for item in y:
    if 'LU' in item:
        df["sample"][i] = 'LUNG'
    elif 'HA' in item:
        df["sample"][i] = 'LYPMH'
    else:
        df = df.drop([i])
    i = i+1
    print(i)
```

Problem definisanja mutiranih gena

- Start: 24,908,712
- End: 249,133,375
- Num_Probes: 72,607
- Rešenje:
  1. SeparateChromosomes.py
  2. DefineGenes.py
  3. MergeChromosomes.py

Problem definisanja mutiranih gena

Listing 2: Eksplicitno Definisanje gena

```
range_index = int(df['Length'].mean())
df = df.sort_values("Start")
df = df.reset_index(drop=True)

# g = 0 for the first file
# for the rest of them, g is appended
g = g+1
df.iloc[0, df.columns.get_loc('depMapID')] = g
start_old = df["Start"][0]
```

Problem definisanja mutiranih gena

Listing 3: Eksplicitno definisanje gena

```
n = len(df.index)
for i in range(1,n):
    start_new = df["Start"][i]
    if start_new != start_old and not
    (start_new < start_old + range_index):
        g = g + 1
        start_old = start_new
    df.iloc[i, df.columns.get_loc('depMapID')]
    = g
```

Problem korelisanih atributa

- Chromosome i **Gene**
- **Start** i End
- **Num_Probes** i Length
- **Modal_HSCN_1** i LOH
- **Modal_HSCN_2** i Modal_Total_CN
- **Cancer_cell_frac_a1**, Ccf_ci95_low_a1 i Ccf_ci95_high_a1
- **Cancer_cell_frac_a2**, Ccf_ci95_low_a2 i Ccf_ci95_high_a2

Drveta odlučivanja

- Mera nečistoće: entropija
- Maksimalna dubina: 5

Tabela: Analiza drveta odlučivanja

	preciznost	**f1-skor**
LUNG	62.00%	71.00%
LYMPH	68.00%	53.00%

Najbliži susedi

- Veličina trening skupa: 95%
- Broj suseda: 9
- Euklidsko rastojanje
- Podjednak uticaj svih suseda

Tabela: Analiza modela k najbližih suseda

	preciznost	**f1-skor**
LUNG	64.00%	65.00%
LYMPH	61.00%	59.00%

Neuronske mreže

- Funkcija aktivacije: tangens hiperbolički
- Veličina skrivenig sloja: (10, 3)
- Stopa učenja: prilagodljiva
- Inicijalna stopa učenja: 4
- Maksimalan broj iteracija: 45

Tabela: Analiza MLP modela

	preciznost	**f1-skor**
LUNG	62.00%	69.00%
LYMPH	66.00%	56.00%

Gausova klasifikacija

Tabela: Analiza modela dobijenog Gausovom klasifikacijom

	preciznost	**f1-skor**	**tačnost**
LUNG	58.00%	60.00%	65.74%
LYMPH	72.00%	70.00%	65.74%

Zaključak

Da li je istraživanje uspešno?

- Prosečna preciznost pogađanja: 63%
- Ne: preciznost veća od 90%
- Da: upoznavanje sa osobinama kancerogenih tkiva
- Naredni koraci:
  - Pravila pridruživanja
  - Eksperimentisati sa različitim kombinacijama klasa
  - PCA amplifikacija
  - Poboljšati funkcionalnost definisanja gena

Hvala na pažnji!