Prepoznavanje vrste kancera

Metoda klasifikacije

N. Bogdanović

Univerzitet u Beogradu:
Matematički Fakultet

25. juni 2019

Plan izlaganja

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 i SPSS Modeler

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

SPSSModeler → Split:

- 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
- SPSSModeler → Append: EXTRACTED_CLASSES.xlsx
- final.xlsx

Problem prevelikog broja klasa

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] = 'HAEMATOPOIETIC'
    else:
        df = df.drop([i])
    i = i+1
    print(i)
```

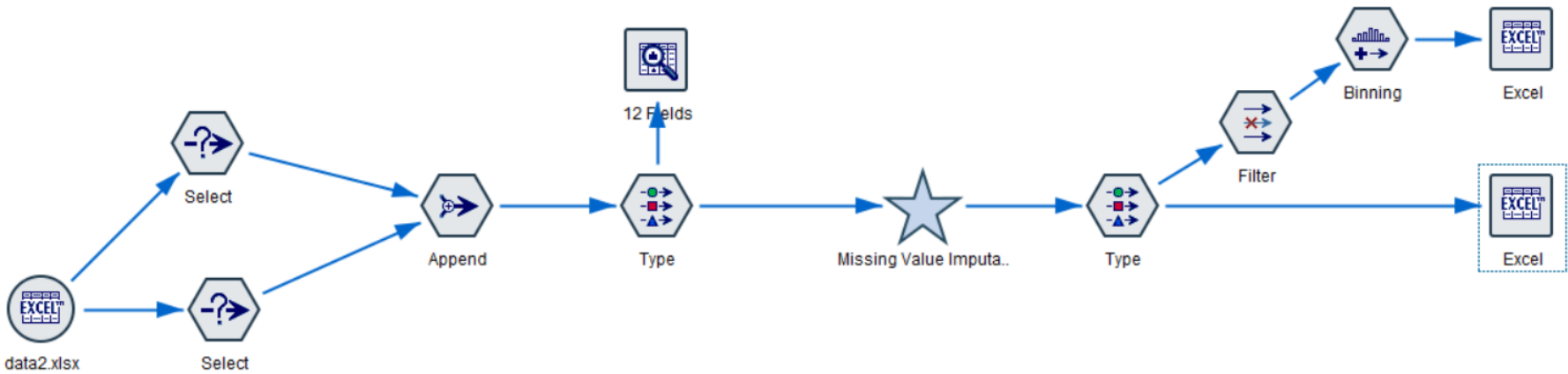
Problem nedostajućih vrednosti

- SPSSModeler → Data Audit → Quality
- 98.47% kompletni podaci
- Zamena srednjim vrednostima

Problem definisanja mutiranih gena

- Start: 24,908,712
- End: 249,133,375
- Num_Probes: 72,607
- Rešenje: implicitno i eksplicitno definisanje gena

Problem definisanja mutiranih gena



Slika: Definisanje gena

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)

pd.options.mode.chained_assignment = None
# 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
```

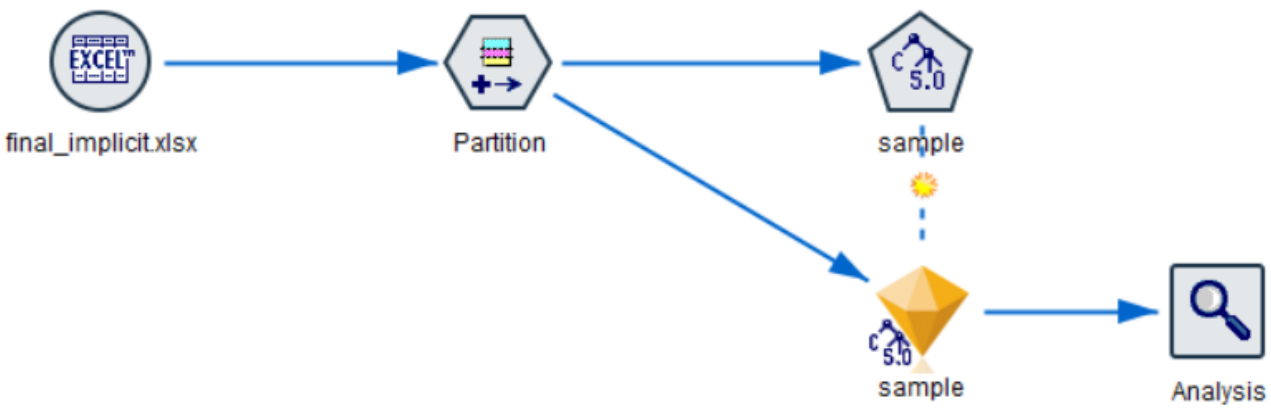
Problem korelisanih atributa

- 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

SPSS

- Simboličko grupisanje
- Poboljšanje produktivnosti (*boosting*)
- Unakrsna validacija



Slika: Drveta odlučivanja - SPSS

Drveta odlučivanja

SPSS - C5.0

Implicitno:

- Dubina drveta: 23
- Sredina 70.5
- Standardna greška: 0.2
- Najbitniji atribut za odlučivanje: Modal_HSCN_2
- Najmanje bitan atribut za odlučivanje: Subclonal_HSCN_a2
- Procenat pogođenih: 72.09%, 71.28%, 70.43%
- Procenat promašenih: 27.91%, 28.72%, 29.57%

Eksplicitno:

- Dubina drveta: 17
- Sredina 65.2
- Standardna greška: 0.2
- Najbitniji atribut za odlučivanje: Modal_HSCN_2
- Najmanje bitan atribut za odlučivanje: Modal_HSCN_1
- Procenat pogođenih: 72.53%, 65.08%, 65.39%
- Procenat promašenih: 27.47%, 34.92%, 34.61%

Drveta odlučivanja

SPSS - CR&T

- Cilj: novi model u vidu drveta odlučivanja
- Maksimalna dubina: 12
- Minimalni broj instanci u grani roditelja: 5%
- Minimalni broj instanci u grani deteta: 2%
- Mera nečistoće: Gini
- Minimalna promena u nečistoći: 0.0001%
- Najbitniji atribut: Modal_HSCN_2
- Najmanje bitan atribut: Subclonal_HSCN_a2
- Dubina drveta: 4
- Tačnost: 69.92%
- Procenat pogođenih: 69.92%, 70.18%, 69.85%
- Procenat promašenih: 30.08%, 29.82%, 30.15%

Drveta odlučivanja

Python

Tabela: Analiza drveta odlučivanja - implicitno grupisanje

	preciznost	**f1-skor**	**tačnost**
LUNG	73.00%	74.00%	68.42%
HAEMATOPOIETIC	63.00%	61.00%	68.42%

Najbliži susedi

SPSS

- Cilj: predviđanje klase
- Analiza: balansirana, brza i tačna
- Ciljno polje: sample
- Minimalni broj suseda: 3
- Maksimalan broj suseda: 5
- Euklidska udaljenost

- Najbolji rezultati: k = 5
- Tačnost: 74.42%
- Procenat pogođenih: 77.28%, 77.15%, 75.00%
- Procenat promašenih: 22.72%, 22.85%, 25.00%

Najbliži susedi

Python

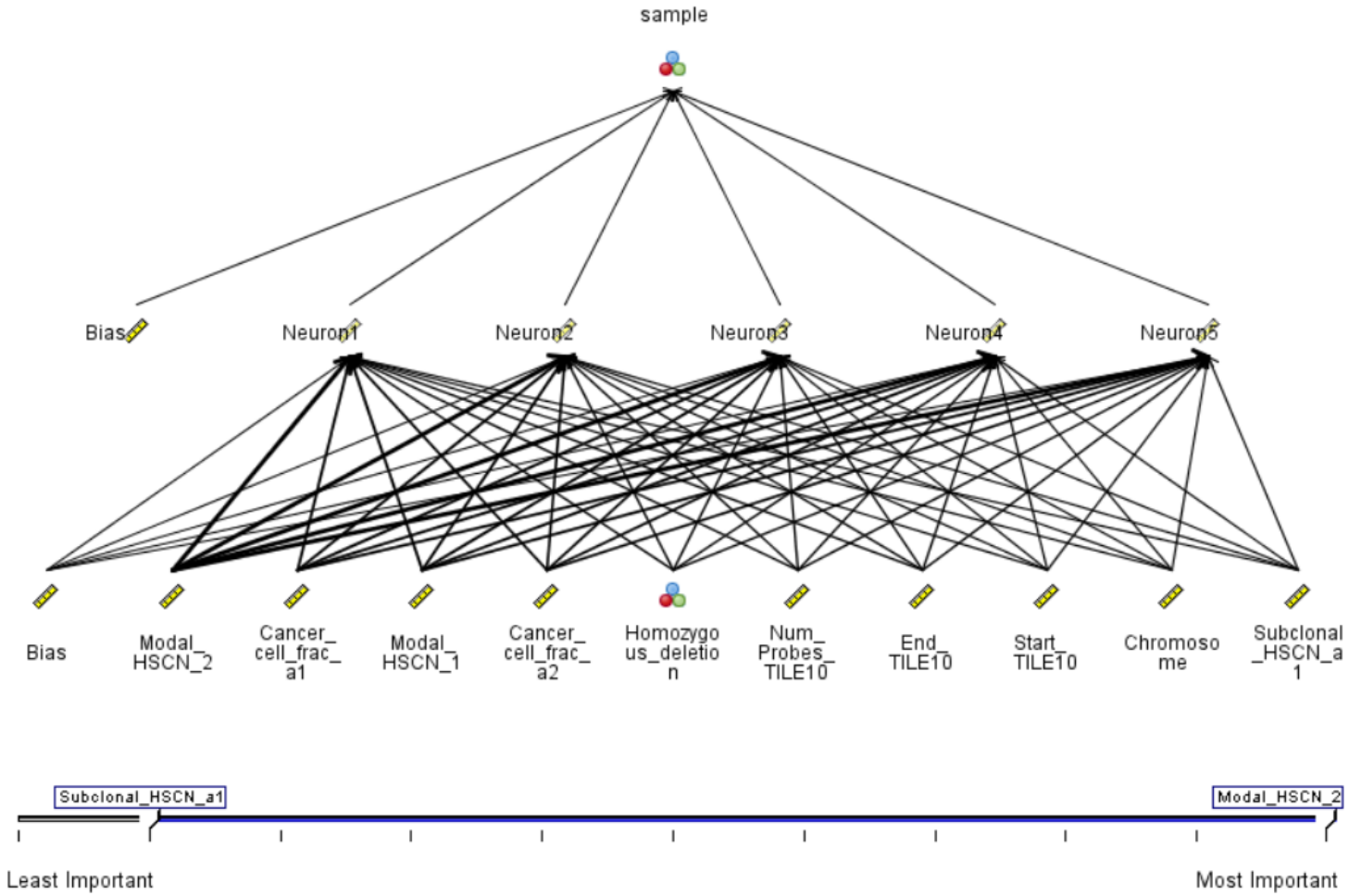
- Veličina trening skupa: 95%
- k = 9
- Euklidsko rastojanje
- Podjednak uticaj svih suseda

Tabela: Analiza modela k najbližih suseda - Python

	preciznost	**f1-skor**	**tačnost**
LUNG	72.00%	74.00%	68.51%
HAEMATOPOIETIC	63.00%	60.00%	68.51%

Neuronske mreže

SPSS



Slika: Neuronska mreža

Neuronske mreže

SPSS

'Partition'	1_Training		2_Testing		3_Validation	
Correct	32,180	69.74%	9,140	70.2%	4,737	69.94%
Wrong	13,964	30.26%	3,880	29.8%	2,036	30.06%
Total	46,144		13,020		6,773	

Slika: Analiza modela neuronske mreže - SPSS

Neuronske mreže

Python

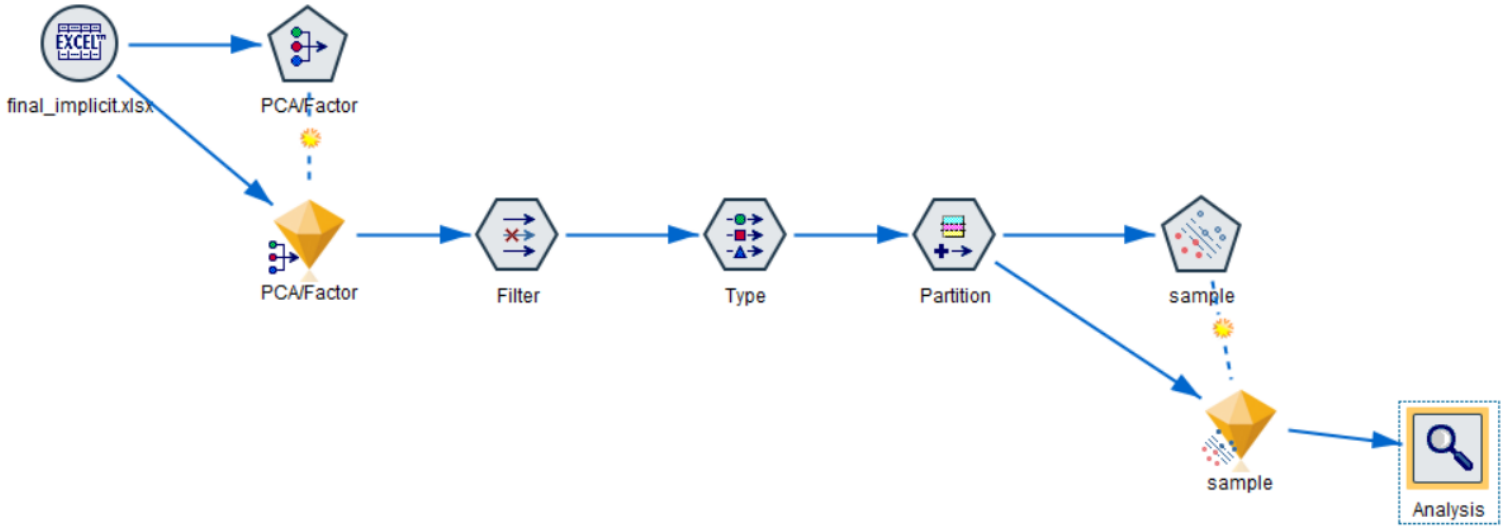
- Funkcija aktivacije: tangens hiperbolički
- Veličina skrivenig sloja: (10, 10)
- Stopa učenja: prilagodljiva
- Inicijalna stopa učenja: 0.01
- Maksimalan broj iteracija: 500

Tabela: Analiza MLP modela - Python

	preciznost	**f1-skor**	**tačnost**
LUNG	71.00%	76.00%	69.74%
HAEMATOPOIETIC	67.00%	59.00%	69.74%

Metod potpornih vektora

SPSS



Slika: Metod potpornih vektora primenjen na 5 najznačajnijih komponenti

'Partition'	1_Training		2_Testing		3_Validation	
Correct	31,323	67.88%	8,843	67.92%	4,598	67.89%
Wrong	14,821	32.12%	4,177	32.08%	2,175	32.11%
Total	46,144		13,020		6,773	

Slika: Analiza modela potpornih vektora

Gausova klasifikacija

Python

Tabela: Analiza modela dobijenog Gausovom klasifikacijom

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

Zaključak

Da li je istraživanje uspešno?

- Najbolji rezultati: Najbliži susedi - 75% tačnosti
- Najlošiji rezultati: Gausova klasifikacija - 65% tačnosti
- Pogled na istraživanje: preciznost veća od 90% i upoznavanje sa osobinama kancerogenih tkiva
- Naredni koraci:
  - Pravila pridruživanja
  - Eksperimentisati sa matricom cene
  - Eksperimentisati sa različitim klasama
  - Eksperimentisati sa eksplicitnim definisanjem gena

Hvala na pažnji!