Komparativna analiza algoritama klasifikacije

Danica Đorđević 1121

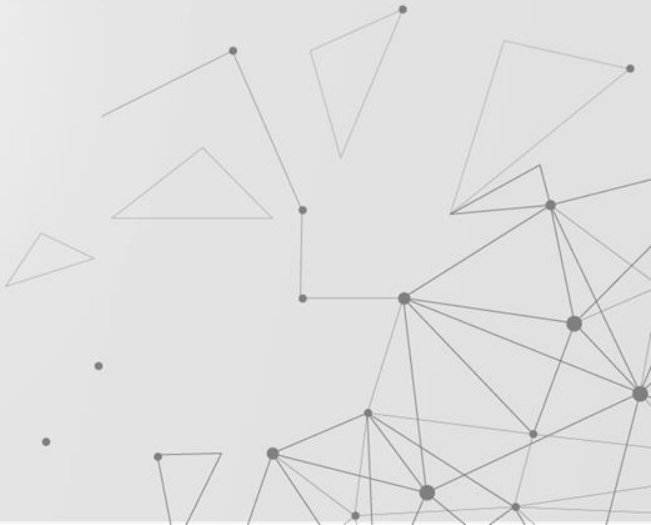
Komparativna analiza algoritama klasifikacije

U ovom radu je vršeno poređenje između sledećih algoritama:

Naivni Bajesov algoritam,

Algoritam K-najbližih suseda,

Algoritam zasnovan na stablima odluke (CART).



Preprocesiranje

```
Atribut 'workclass' ima 8 jedinstvenih vrednosti
Atribut 'education' ima 16 jedinstvenih vrednosti
Atribut 'marital_status' ima 7 jedinstvenih vrednosti
Atribut 'occupation' ima 14 jedinstvenih vrednosti
Atribut 'relationship' ima 6 jedinstvenih vrednosti
Atribut 'race' ima 5 jedinstvenih vrednosti
Atribut 'sex' ima 2 jedinstvenih vrednosti
Atribut 'native_country' ima 40 jedinstvenih vrednosti
```

Potencijalni kandidati za smanjenje broja jedinstvenih vrednosti su kolone *native_country*, *education* i *occupation*.

```
United-States    5545
Mexico            138
?                 108
Philippines        32
Canada             28
Germany            28
Puerto-Rico        23
Jamaica            22
China              18
South              18
Name: native_country, dtype: int64
```

U zavisnostiod distribucija vrednosti, može se izvršiti smanjenje broja jedinstvenih vrednosti za atribut *native_country* jer 90% ispitanika dolazi iz Amerike. Nove vrednosti za stribut su *United-States* i *Other*.

```
Prof-specialty     793
Exec-managerial    755
Adm-clerical       751
Craft-repair       741
Sales              717
Other-service      627
Machine-op-inspct  384
?                  357
Transport-moving   311
Handlers-cleaners  255
Name: occupation, dtype: int64
```

```
HS-grad         2007
Some-college    1451
Bachelors        997
Masters          311
Assoc-voc        263
Assoc-acdm       205
11th             202
10th             187
7th-8th          134
Prof-school      112
Name: education, dtype: int64
```

Za atribute *education* i *occupation* se ne može izvršiti smanjenje broja jedinstvenih vrednosti jer ne postoji vrednost atributa koja predstavlja većinu.

Preprocesiranje

```
 <=50K    4721
 >50K     1483
Name: income, dtype: int64
```

Klasni atribut prevesti u numericki oblik, gde >50K predstavlja 1, a <=50K predstavlja 0

```
def _replace_outcome_values(y):
    for i in range(len(y)):
        val = y[i].strip()
        if val == ">50K":
            y[i] = 1
        else:
            y[i] = 0
    return y
```

Funkcija za zamenu vrednosti klasnog atributa

Kako modeli rade samo sa numeričim vrednostima, neophodno je ove atribute prevesti iz nenumeričkih u numeričke oblike. Biblioteka pandas nudi rešenje za ovaj problem, korišćenjem dummies atributa. Ovaj metod za svaku jedinstvenu vrednost koju kolona poseduje, pravi nove kolone, a staru kolonu briše. Tako, na primer, ako se primeni ova metoda na kolonu race, dobiće se 5 novih kolona sa binarnim vrednostima, dok će stara race kolona biti izbrisana.

	Amer-Indian-Eskimo	Asian-Pac-Islander	Black	Other	White
0	0	0	0	0	1
1	0	0	0	0	1
2	0	0	0	0	1
3	0	0	0	0	1
4	0	0	0	0	1

Preprocesiranje

```
def _impute_missing_values(X):
    si = SimpleImputer(missing_values=np.nan, strategy='mean')
    si.fit(X)
    X = pd.DataFrame(data=si.transform(X), columns=X.columns)
    return X
```

Zamena nedostajućih vrednosti srednjom vrednošću

Dodavanje interakcija

	age	fnlwgt	education_num	capital_gain	...	race_ White_sex_ Male	race_ White_native_country_Other	sex_ Female_native_country_Other	sex_ Male_native_country_Other
0	38.0	234962.0	9.0	2829.0	...	1.0	1.0	0.0	1.0
1	72.0	177226.0	9.0	0.0	...	1.0	1.0	0.0	1.0
2	31.0	259931.0	4.0	0.0	...	1.0	1.0	0.0	1.0
3	55.0	189528.0	4.0	0.0	...	1.0	1.0	0.0	1.0
4	38.0	34996.0	10.0	0.0	...	0.0	1.0	1.0	0.0

[5 rows x 1630 columns]

Preprocesiranje

```
'------------------------- TREE --------------------'
'AUC modela sa preprocesiranjem: 0.7283756413718606'
'AUC modela bez preprocesiranja: 0.6694099378881988'
'Poboljsanje modela: 8.808608917531478%'
'---------------------------------------------------'
('------------------------- BAYES Multinominal '
 '----------------------------------')
'AUC modela sa preprocesiranjem: 0.3620375371320551'
'AUC modela bez preprocesiranja: 0.5723737510126924'
'Poboljsanje modela: -36.74805378627034%'
'---------------------------------------------------'
'------------------------- K NEIGHBORS -------------'
'AUC modela sa preprocesiranjem: 0.8094652984066972'
'AUC modela bez preprocesiranja: 0.5988320280853363'
'Poboljsanje modela: 35.17401549059175%'
'---------------------------------------------------'
'------------------------- TREE PCA ----------------'
'AUC modela sa preprocesiranjem: 0.7408385093167702'
'AUC modela bez preprocesiranja: 0.6777950310559007'
'Poboljsanje modela: 9.301260022909503%'
'---------------------------------------------------'
('------------------------- BAYES Multinominal PCA '
 '----------------------------------')
'AUC modela sa preprocesiranjem: 0.833918444504456'
'AUC modela bez preprocesiranja: 0.6021874156089656'
'Poboljsanje modela: 38.48154625767976%'
'---------------------------------------------------'
'------------------------- K NEIGHBORS PCA ---------'
'AUC modela sa preprocesiranjem: 0.6005941128814475'
'AUC modela bez preprocesiranja: 0.5988320280853363'
'Poboljsanje modela: 0.2942535992513862%'
'---------------------------------------------------'
```

Često ulazne karakteristike interaguju na neočekivan i nelineralni način prilikom prediktivnog modeliranja. Te interakcije se mogu identifikovati i modelirati pomoću algoritma za učenje. Drugi pristup je osmišljavanje novih karakteristika koje uočavaju ove interakcije i utvrđivanje da li poboljšavaju performanse modela.

Performanse klasifikatora su znatno gore, ako se ne obavi korak dodavanja interakcija.

Preprocesiranje – Redukcija dimanzionalnosti

```python
def _pca(X):
    pca = PCA(n_components=10)
    X_pca = pd.DataFrame(pca.fit_transform(X))
    return X_pca
```

Korišćenje PCA metode

	0	1	2	3	4	5	6	7	8	9
0	4.603680e+08	-1.565683e+07	7.088114e+05	1.704761e+06	1.473788e+05	-52838.823539	233399.450121	153546.197217	167612.864361	-86417.009491
1	-2.043396e+08	-1.651429e+07	-5.935961e+05	8.347809e+06	1.072941e+05	-685.779770	162372.075717	86929.540199	-19285.642779	-80394.262033
2	-2.043383e+08	-1.648310e+07	2.642779e+06	-1.177344e+06	-1.268122e+06	-27744.038930	209406.504532	124026.132287	48292.490478	96700.176617
3	-2.043374e+08	-1.646142e+07	4.896327e+06	-9.190717e+04	-1.858992e+06	-32092.377953	93546.539915	42950.529469	-38964.830372	68647.632896
4	-2.043425e+08	-1.659134e+07	-8.516450e+06	1.705205e+04	-2.587862e+05	-45291.892752	1504.131647	-54915.983056	3047.100899	24799.014780
...	...	...	...	...	...	...	...	...	...	...
6199	-2.043395e+08	-1.651589e+07	-7.595636e+05	-3.905519e+06	-2.964177e+05	-35100.408105	-190262.990397	-66513.436771	128721.817277	-60992.357319
6200	-2.043389e+08	-1.649903e+07	9.907504e+05	2.198097e+06	-4.030509e+05	-29237.665461	-221283.534397	-33696.082799	74679.528787	-74820.967918
6201	-2.043352e+08	-1.640423e+07	1.075960e+07	2.442868e+06	-9.611259e+05	-12889.197566	-225086.311551	177070.498337	-209992.013130	-88907.992400
6202	-2.043399e+08	-1.652669e+07	-1.881780e+06	-2.694645e+06	1.389111e+05	-33471.930392	-218206.516380	6130.424226	65592.354062	-130637.564220
6203	-2.043373e+08	-1.645879e+07	5.129190e+06	5.330803e+04	-5.904585e+05	-20196.586251	-116003.375779	176838.162313	-13040.662382	-122795.358481

Preprocesiranje – Redukcija dimanzionalnosti

Selekcija K najrelevantnijih atributa

```
select = sklearn.feature_selection.SelectKBest(f_classif,k=16)
selected_features = select.fit(X_train, y_train)
indices_selected = selected_features.get_support(indices=True)
colnames_selected = [X.columns[i] for i in indices_selected]

X_train_selected = X_train[colnames_selected]
X_test_selected = X_test[colnames_selected]
```

```
'Selected columns: 16'
['marital_status_ Married-civ-spouse',
 'relationship_ Husband',
 'age_education_num',
 'age_marital_status_ Married-civ-spouse',
 'age_relationship_ Husband',
 'education_num_marital_status_ Married-civ-spouse',
 'education_num_relationship_ Husband',
 'hours_per_week_marital_status_ Married-civ-spouse',
 'hours_per_week_relationship_ Husband',
 'marital_status_ Married-civ-spouse_relationship_ Husband',
 'marital_status_ Married-civ-spouse_race_ White',
 'marital_status_ Married-civ-spouse_sex_ Male',
 'marital_status_ Married-civ-spouse_native_country_Other',
 'relationship_ Husband_race_ White',
 'relationship_ Husband_sex_ Male',
 'relationship_ Husband_native_country_Other']
```

Preprocesiranje – Redukcija dimanzionalnosti

Selekcija K najrelevantnijih atributa

```
select = sklearn.feature_selection.SelectKBest(f_classif,k=16)
selected_features = select.fit(X_train, y_train)
indices_selected = selected_features.get_support(indices=True)
colnames_selected = [X.columns[i] for i in indices_selected]

X_train_selected = X_train[colnames_selected]
X_test_selected = X_test[colnames_selected]
```

```
'Selected columns: 16'
['marital_status_ Married-civ-spouse',
 'relationship_ Husband',
 'age_education_num',
 'age_marital_status_ Married-civ-spouse',
 'age_relationship_ Husband',
 'education_num_marital_status_ Married-civ-spouse',
 'education_num_relationship_ Husband',
 'hours_per_week_marital_status_ Married-civ-spouse',
 'hours_per_week_relationship_ Husband',
 'marital_status_ Married-civ-spouse_relationship_ Husband',
 'marital_status_ Married-civ-spouse_race_ White',
 'marital_status_ Married-civ-spouse_sex_ Male',
 'marital_status_ Married-civ-spouse_native_country_Other',
 'relationship_ Husband_race_ White',
 'relationship_ Husband_sex_ Male',
 'relationship_ Husband_native_country_Other']
```

Preprocesiranje

```
X_train, X_test, y_train, y_test = train_test_split(
    X, y, train_size=0.9, random_state=1)
```

Kros validacija

Priprema neobrađenog skupa podataka, na osnovu koga će se vršiti poređenje. Iz skupa neprocesiranih podataka su izbačene kolone sa nenumeričkim podacima, kao i redovi sa nedostajućim vrednostima. Nenumeričke podatke je neophodno izbaciti, da bi primenjivanje bilo kog algoritma bilo moguće.

	age	fnlwgt	education_num	capital_gain	capital_loss	hours_per_week
0	38	234962	9	2829	0	30
1	72	177226	9	0	0	8
2	31	259931	4	0	0	40
3	55	189528	4	0	0	60
4	38	34996	10	0	0	40

Algoritmi klasifikacije – sklearn biblioteka

```python
def _find_model_performance_naive_bayes(X_train, y_train, X_test, y_test):
    classifier = MultinomialNB()
    classifier.fit(X_train, y_train)
    y_hat = [x[1] for x in classifier.predict_proba(X_test)]
    auc = roc_auc_score(y_test, y_hat)
    return auc


def _find_model_performance_tree(X_train, y_train, X_test, y_test):
    classifier = DecisionTreeClassifier()
    classifier.fit(X_train, y_train)
    y_hat = [x[1] for x in classifier.predict_proba(X_test)]
    auc = roc_auc_score(y_test, y_hat)

    return auc


def _find_model_performance_k_neighbors(X_train, y_train, X_test, y_test):
    classifier = KNeighborsClassifier()
    classifier.fit(X_train, y_train)
    y_hat = [x[1] for x in classifier.predict_proba(X_test)]
    auc = roc_auc_score(y_test, y_hat)

    return auc
```

Komparativna analiza algoritama

Selekcija K najrelevantnijih atributa

```
'------------------------ TREE --------------------'
'AUC modela sa preprocesiranjem: 0.7107953011072105'
'AUC modela bez preprocesiranja: 0.6613354037267081'
'Poboljsanje modela: 7.478791714731988%'
'--------------------------------------------------'
('------------------------ BAYES Multinominal '
 '---------------------------------')
'AUC modela sa preprocesiranjem: 0.7244733999459898'
'AUC modela bez preprocesiranja: 0.5723737510126924'
'Poboljsanje modela: 26.57348431233783%'
'--------------------------------------------------'
'------------------------ K NEIGHBORS -------------'
'AUC modela sa preprocesiranjem: 0.8059006211180123'
'AUC modela bez preprocesiranja: 0.5988320280853363'
'Poboljsanje modela: 34.57874384153143%'
'--------------------------------------------------'
```

PCA metoda

```
'------------------------ TREE PCA ----------------'
'AUC modela sa preprocesiranjem: 0.7104037267080745'
'AUC modela bez preprocesiranja: 0.6821428571428572'
'Poboljsanje modela: 4.142954700660133%'
'--------------------------------------------------'
('------------------------ BAYES Multinominal PCA '
 '---------------------------------')
'AUC modela sa preprocesiranjem: 0.7812314339724549'
'AUC modela bez preprocesiranja: 0.6021874156089656'
'Poboljsanje modela: 29.7322749899099%'
'--------------------------------------------------'
'------------------------ K NEIGHBORS PCA ----------'
'AUC modela sa preprocesiranjem: 0.7480556305698083'
'AUC modela bez preprocesiranja: 0.5988320280853363'
'Poboljsanje modela: 24.919108445416505%'
'--------------------------------------------------'
```

Evaluacija algoritama

Za evaluaciju algoritama je korišćena AUC (Area Under the ROC Curve) mera. AUC meri celokupno dvodimenzionalno područje ispod cele ROC krive od (0,0) do (1,1). . Model čija su predviđanja 100% pogrešna ima AUC od 0,0. Onaj model čija su predviđanja 100% tačna ima AUC 1,0.

```
def _find_model_performance_k_neighbors(X_train, y_train, X_test, y_test):
    classifier = KNeighborsClassifier()
    classifier.fit(X_train, y_train)
    y_hat = [x[1] for x in classifier.predict_proba(X_test)]
    auc = roc_auc_score(y_test, y_hat)

    return auc
```

Korićena je *roc_auc_score* funkcija iz *sklearn* biblioteke

```
performanse_improve = (
    (performanse_processed-performanse_unprocessed)/performanse_unprocessed)*100
```

Računjanje procenta poboljšanja klasifikatora

Zaključak

Hvala na pažnji.