Analiza '101 Innovations - Research Tools Survey' skupa podataka

Seminarski rad u okviru kursa
Istraživanje podataka 1
Matematički fakultet

Dimitrije Antić
mi16128@alas.matf.bg.ac.rs

avgust 2019.

Sažetak

U ovom radu su dati rezultati istraživanja skupa podataka *101 Innovations - Research Tools Survey*. Nakon kratkog opisa strukture samog skupa, opisan je nacin na koji je on obrađen kako bi se prilagodio korišćenim alatima. Veliki broj informacija je dobijen kao izlaz iz određenih algoritama. Postupak analize podataka je opisan u daljem tekstu, a najbolje rezultate pri klasifikaciji je dala neuronska mreža.

Sadržaj

1 Opis skupa podataka

101 Innovations - Research Tools Survey je skup podataka koji sadrži informacije o online anketama koje su popunjavali korisnici, odgovarajući na pitanja o 17 istraživačkih aktivnosti. 20,663 ankete su popunjene u periodu od 10.5.2015. do 10.2.2016.

Sadrži 178 kolona, i 20663 instanci. Svaka od kolona je kategoričkog tipa, skup podataka ne sadrži ni jedan neprekidni atribut. Veliki broj kolona je binaran i predstavlja jedan od ponuđenih odgovora unapred zadatih u anketi.

Takođe, u svakoj grupi ponuđenih odgovora postoji i odgovor *eng. Other* koje ispitaniku daje mogućnost popunjavanja tzv. *eng. open text* polja odnosno kolona u kojima ispitanik sam daje odgovor.

Ciljni atribut je atribut **eng. ROLE** koji predstavlja ulogu ispitanika, i inicijalno sadrži 8 različitih vrednosti.

Skup je u svojoj originalnoj formi organizovan u CSV format, u istoj formi je i korišćen u istraživanju. Oblik podataka je menjan u zavisnosti od korišćenog alata odnosno algoritma.

2 Korišćeni alati

Za obradu i analizu podataka korišćen je alat *SPSS* i programski jezik *Python* i pripadajuće biblioteke poput *sklearn*, *numpy*, *pandas*, *matplotlib*, *keras* odnosno *TensorFlow*.

Korišćena Jupyter sveska kao i *tok podataka* korišćen u SPSS-u priloženi su uz rad.

3 Preprocesiranje

U ovom poglavlju ce biti objašnjen proces preprocesiranja, odnosno kako je skup podataka pripremljen za algoritme klasifikacije.

Sam proces preprocesiranja vršen je u više koraka.

Jedan od koraka u preprocesiranju je predstavljao prevođenje odgovora iz tekstualnog oblika u broj, kako bi neuronska mreža na ulazu dobila vektore.

Odgovori su dati kao odvojeni kolone, ali mogu se organizovati u 17 grupa. Svaka od grupa predstavlja odgovor na jedno pitanje u vezi sa istraživanjima ispitanika. Tabela je data na sledećoj strani.

Tabela 1: Kolone u skupu podataka

Grupa atributa	**Opis grupe atributa**
ROLE	ciljni atribut, uloga istraživača
ROLESPECCL	open text odgovor za ulogu istraživača
COUNTRYCL	open text odgovor za državu stanovanja
group_1	disciplina (nauka) kojom se bavi istraživač
PUBYEAR	interval godina prvog objavljenog rada
group_2	alat/sajt korišćen za pretragu literature
group_3	alat/sajt korišćen za pristup literaturi
group_4	alat/sajt korišćen za predloge literature
group_5	alat/sajt korišćen za pregledanje literature
group_6	alat/sajt korišćen za analiziranje podataka
group_7	alat/sajt korišćen za deljenje radova/protokola
group_8	alat/sajt korišćen za pripremu/pisanje radova
group_9	alat/sajt korišćen za organizaciju referenci
group_10	alat/sajt korišćen za arhiviranje radova
group_11	alat/sajt korišćen za arhiviranje/deljenje programskih kodova
group_12	alat/sajt korišćen za izbor časopisa za objavu rada
group_13	alat/sajt korišćen za objavljivanje radova
group_14	alat/sajt korišćen za edukaciju van akademskih aktivnosti
group_15	korišćeni istraživački profili
group_16	alat/sajt korišćen za pregledanje/recenziju radova
group_17	alat/sajt korišćen za merenje uticaja

Kako je nezanemarljiv broj ispitanika popunjavao *eng. open text* polja, odlučeno je da se iz svakog od tih polja najbrojnija vrednost izvuče kao još jedna vrednost odnosno kolona zarad dobijanja dodatnih informacija. Kriterijum koji je utvrđen eksperimentalno: ako je broj pojavljivanja najbrojnije vrednosti veći od 10% broja ljudi koji su popunjavali tu kolonu, vrednost je izvučena u suprotnom nije.

Jedna od kolona bila je *PUBYEAR* i kao što je opisano u 3, i bila je očigledni kandidat za prebacivanje u redni tip podataka, s obzirom da su intervali imali hronološku zavisnost. To se ekperimentalno i pokazalo kao dobro rešenje.

Ceo proces preprocesiranja je rađen u programskom jeziku *Python*.

U zavisnosti od algoritma i alata koji je korišćen, skup podataka je transformisan u ciljni oblik.

Oblik podataka korišćen za rad je s obzirom na broj kolona prikazan u jupyter svesci, i predstavljao je:

1. Za rad u alatu SPSS, skup podataka je sadržao kolone navedene u 3, gde je svaka od kolona imala ceo broj različitih vrednosti odnosno kategorija.
2. Za rad u jeziku Python, skup podataka je sadržao 309 kolona (retka matrica), koje su bile binarnog tipa, odnosno predstavljale su 1/NaN polja i davale odgovor da li je ispitanik potvrdio taj unapred zadati odgovor.

4 Klasifikacija

Potrebno je klasifikovani istraživača u 8 inicijalnih klasa. Kako je navedeno u sekciji 3, dodata je još jedna, deveta, kao posledica velikog broja popunjenih polja otvorenog teksta. Za klasifikaciju datog skupa podataka korišćeni su, pored navedenih, algoritmi KNN, logistička regresija iz navedenih alata ali s obzirom na lošije rezultate neće biti prikazivani. Oni koji su prikazali najbolje rezultate dati su u nastavku:

- Neuronska mreža u Python-u
- XGBoost u SPSS-u
- C5.0 algoritam u SPSS-u

Najbolja preciznost je postignuta klasifikacijom korišćenjem neuronskih mreža, i iznosi 0.6224.

4.1 Neuronske mreže

Kako je skup podataka sam po sebi velike dimenzionalnosti, kod neuronskih mreža su dodavane određene modifikacije kako bi na što efikasniji način bilo sprečeno preprilagođavanje modela. Korišćena su unapređenja kao npr. *eng. dropout rate* koji se zadaje verovatnoća sa kojim će se izlaz iz jednog neurona koristiti, odnosno verovatnoća sa kojom se rezultati neurona prihvataju. Kao funkcija greške je korišćena srednje kvadratna greška.

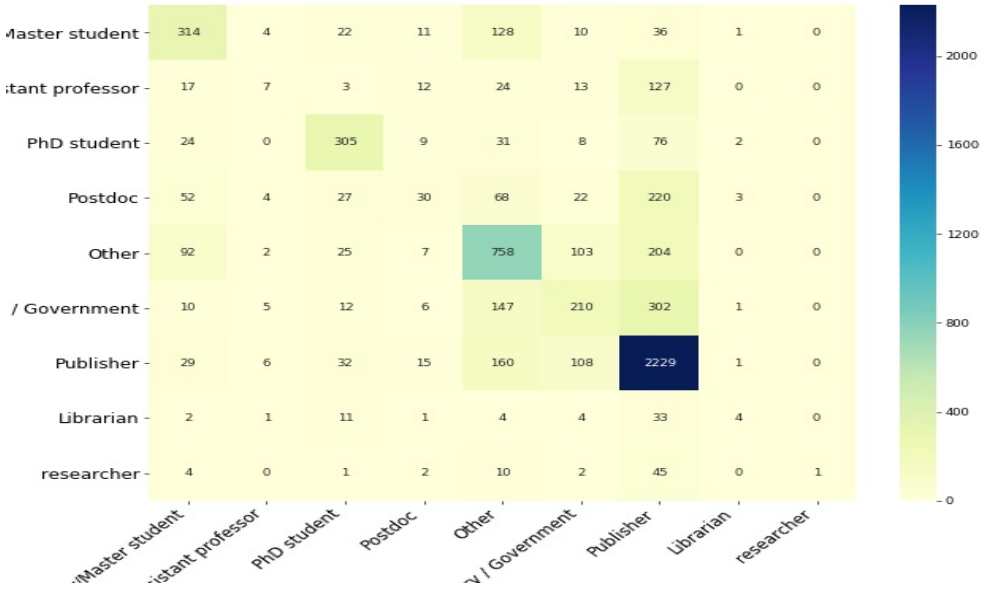
Uz neuronsku mrežu iz biblioteke *Keras*, korišćena je i *eng. auto encoder* neuronska mreža koja je imala za zadatak smanjivanje dimenzionalnosti.

Eksperimenti su vršeni na više različitih oblika podataka dobijenih na načine opisane u u poglavlju 3:

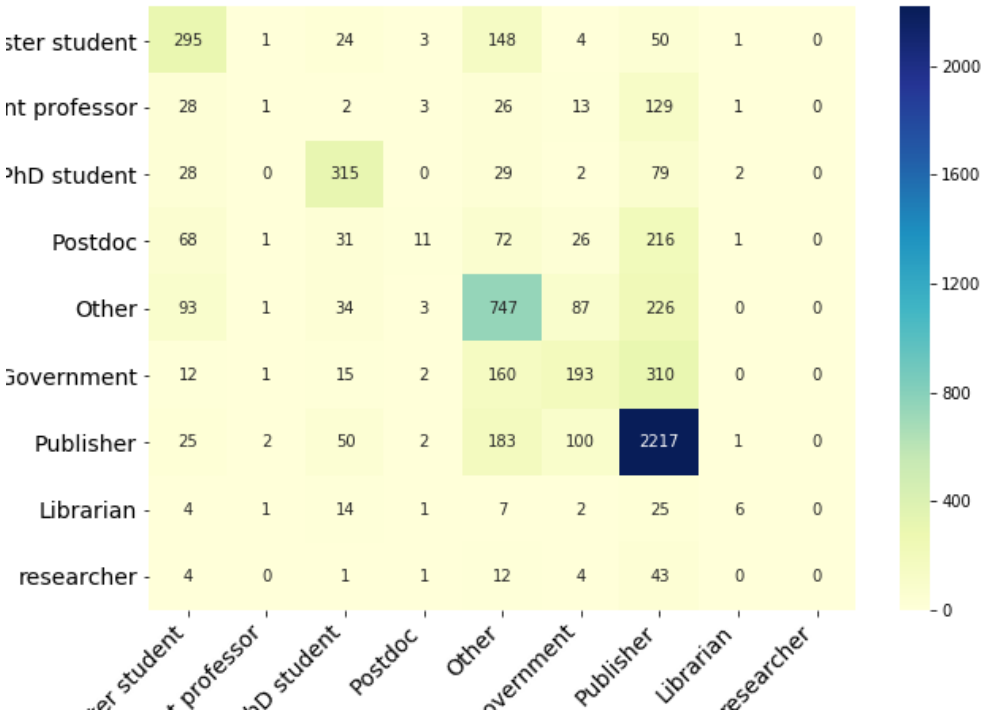
1. Izvučene su nove kolone na osnovu slobodnih odgovora korisnika,
2. Dodata nova klasa, u kombinaciji sa prethodnim oblikom,
3. Nove kolone, nova klasa, sa atributom koji se odnosi na državu istraživača,
4. Nove kolone, nova klasa, bez atributa koji se odnosi na državu istraživača.

Najbolji rezultat je dobijen pri radu sa skupom podataka u obliku **3**, i to sledeće vrednosti:

- **Bez autoencodera**: preciznost: 0.6224, greška: 1.2568
- **Sa autoencoderom**: preciznost: 0.6159, greška: 1.1670



Matrica konfuzije neuronske mreže bez autoencodera



Matrica konfuzije neuronske mreže sa autoencoderom

Iz matrica konfuzije mogu se izvesti zaključci koji se pojavljuju i kod drugih modela, zbog toga biće opisani samo jednom. Neuronska mreža je najpreciznije klasifikovala klasu Publisher, što je i očekivano s obzirom da je najbrojnija, pa je model bio dobro istreniran da je prepozna. Klasa Publisher se ujedno najviše meša i sa Industry/Government, Postdoc, Other, i Asisstant professor, što ima smisla jer je očekivano da su svi iz navedenih krugova sličniji Publisher-u nego npr Master student. Nova klasa nije doprinela preciznosti, ali potvrđuje prethodno navedeni zaključak, vrlo je očekivano od nekoga ko je researcher da ujedno i objavljuje radove.

4.2 XGBoost

Za rad sa ovim algoritmom korišćen je alat *SPSS* i odgovarajući oblik dat u poglavlju 3. Predstavlja varijaciju Random forest ansambla koji koristi *eng. gradient boosting* kako bi se još efikasnije sprečilo preprilagođavanje.

Rezultati klasifikacije korišćenjem *XGBoost* klasifikatora su:

'Partition'	1_Training		2_Testing	
Correct	8,093	56.05%	3,444	55.83%
Wrong	6,346	43.95%	2,725	44.17%
Total	14,439		6,169	

Preciznost algoritma

S obzirom na broj klasa i format matrice konfuzije dobijene iz alata SPSS, matrica neće biti prikazana u radu (zbog nemogućnosti prikazivanja na širini strane), i biće prikaza na odbrani rada. Može se pogledati u *eng. streamu* izvedenom iz SPSS-a.

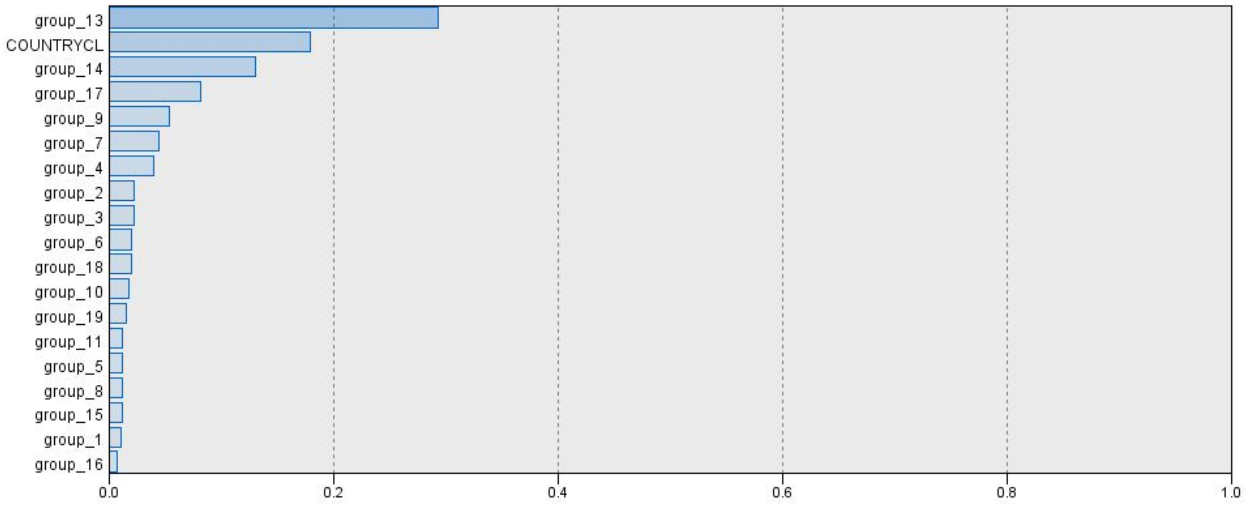
4.3 C5.0

Za rad sa ovim algoritmom korišćen je alat *SPSS* i odgovarajući oblik dat u poglavlju 3. Korišćen je algoritam C5.0 sa *eng. boosting*-om, i ostvario sledeće rezultate:

'Partition'	1_Training		2_Testing	
Correct	8,581	59.29%	2,808	45.36%
Wrong	5,891	40.71%	3,383	54.64%
Total	14,472		6,191	

Preciznost na trening i test skupu

S obzirom na broj klasa i format matrice konfuzije dobijene iz alata SPSS, matrica neće biti prikazana u radu (zbog nemogućnosti prikazivanja na širini strane), i biće prikaza na odbrani rada. Može se pogledati u *eng. streamu* izvedenom iz SPSS-a.



Bitnost atributa korišćenih u klasifikaciji

Prehodni grafikon prikazuje bitnost atributa u procesu klasifikacije. Kao najbitniji se pokazao *group_ 13* koji prema opisu iz 3 govori o alatu koji se koristi za objavljivanje radova. Kako postoje osobe koje nisu istraživači koji objavljuju radove, za vrednost ovog polja su odgovarali sa *Other*, a specijalno polje popunjavali vrednošću *eng. none*. Pritom, ta vrednost nije izvedena kao nova kolona jer se ne uklapa u kriterijum naveden u poglavlju 3.

Kao najmanje bitni:

- *group_ 1* koji se odnosi na granu nauke u kojoj se istražuje, što je i očekivano jer se na osnovu nauke ne može jasno zaključiti da li je neko student, istraživač, profesor ili nešto drugo od navedenih klasa.
- *group_ 13* koji predstavlja alat koji se koristi za recenziju radova.

5 Zaključak

Dobijeni rezultati se oslanjaju na podskup skupa podataka uz dodate nove kolone koje su doprinele preciznosti klasifikatora i dodatu novu klasu.

Skup podataka nije standardizovan, odnosno dajući ispitaniku mogućnost da popuni polja otvorenog odgovora gubi se na konzistentnosti tj. povećava se šansa da odgovori čine nepotpunu informaciju. Samim tim, proces preprocesiranja i klasifikacije podataka je postaje zahtevniji. Veliki problem je bila dimenzionalnost koja s obzirom na tip podataka nije mogla biti redukovana bez gubitka informacija, a doprinosila je preprilagođavanju modela. Dodatna preciznost bi možda bila postignuta detaljnijom obradom polja otvorenih odgovora.