Procenjivanje hemijskih i fizičkih svojstava zemljišta na osnovu spektralnih merenja

Dimitrije Milenković
Fakultet organizacionih nauka, Univerzitet u Beogradu
dm20193047@student.fon.bg.ac.rs

1. Uvod

1.1. Opis problema

Procenjivanje kvaliteta zemljišta u kontekstu funkcionalnih svojstava je od velike važnosti za čoveka tokom čitavog njegovog razvoja. Funkcionalna svojstva zemljišta su ona svojstva koja se odnose na sposobnost tla da podrži osnovne potrebe jednog ekosistema kao što su primarna produktivnost, zadržavanje hranljivih sastojaka i vode i otpornost na eroziju tla. Na osnovu ovih svojstava može se proceniti koliko je određeno zemljište poljoprivredni pogodno i planirati upravljanje njim. Napredak u brzoj i jeftinoj analizi uzoraka tla korišćenjem infracrvene spektroskopije, georeferenciranjem uzoraka tla i većom dostupnošću podataka sa daljinskih senzora zemlje pružaju nove mogućnosti za predviđanje funkcionalnih svojstava. Daljinsko procenjivanje funkcionalnih svojstava tla, posebno u retko naseljenim regionima kao što je Afrika, od izuzetne je važnosti za planiranje održivog poljoprivrednog delovanja i upravljanja prirodnim resursima.
Difuzna reflektivna infracrvena spektroskopija pokazala je potencijal u brojnim istraživanjima da pruži tačna, brza i jeftina merenja mnogih funkcionalnih svojstava tla. Uz minimalnu pripremu uzorka tla, količina svetlosti koju on apsorbuje meri se na veliki broj različitih talasnih opsega u rasponu talasnih dužina da bi se dobio infracrveni spektar. Merenje se obično može izvesti za oko 30 sekundi, za razliku od klasičnih testova koji su spori, skupi i koriste hemikalije.
Funkcionalna svojstva zemljišta mogu se proceniti na osnovu merenja prisustva ključnih materija u njemu. Ovaj projekat predstavlja pristup procenjivanju vrednosti funkcionalnih svojstava na osnovu spektralnih merenja zračenja i geo podataka. Funkcionalna svojstva su predstavljena udelom određene materije u zemljištu, tako da je u pitanju problem višeciljne regresije.

1.2. Opis podataka

Radi se o visoko dimenzionalnom skupu podataka sa 3599 atributa. Nasuprot tome, broj ospervacija nije veliki - ukupno ima 1157 opservacija. U prostoru atributa prediktora nalaze se dva grupe atributa:

- izmerene karakteristike apsorpcije svetlosti:
    - Depth - jedini kategorički atribut, binarni. Označava da li je uzorak uzet iz površinskog ili podzemnog sloja zemlje.
    - m7497.96 - m599.76 - mere upijanja infracrvenog zračenja talasnog broja od 7497.96 do 599.76. Preporuka izvora podataka je da se ukloni spektar CO_2 m2379.76-m2352.76, što je kasnije testirano.
- prostorni atributi:
    - BSA - blizina infracrvenog svetla
    - CTI - topografski indeks izračunat iz nadmorske visine
    - ELEV - nadmorska visina radarske topografije
    - EVI - prosek poboljšanja vegetacionog indeksa
    - LST - prosečne temerature kopna: LSTD za dnevnu, LSTN za noćnu

- Ref - prosek refleksije kopna (Ref1=plava, Ref2=crvena, Ref3=blizu infracrvenog, Ref7=infracrveno)
- Reli - topografski reljef
- TMAP i TMFI - pokazatelji praćenja tropskih šuma (TMAP=srednja godišnja količina padavina i TMFI=modifikovani *Fournier* indeks)

Svi prostorni atributi su skalirani, dok se izmerene karakteristike nalaze na skali (-0.068, 2.569) i, iako mereni na istoj skali, moguće je da bi njihovo skaliranje pomoglo tačnosti, što će kasnije biti testirano.
Skup ima jednu binarnu varijablu *Depth*, dok su sve ostale numeričke. Ne sadrži nedostajuće podatke. Kod većine atributa postoje značajne razlike percentila od minimalne i maksimalne vrednosti, pa je moguće da neke kolone sadrže izuzetke.

	BSAN	BSAS	BSAV	CTI	ELEV	EVI	LSTD	LSTN	REF1	REF2	REF3	REF7	RELI	TMAP	TMFI
mean	-0.57	-0.62	-0.69	-0.21	0.53	0.7	-0.41	-0.09	-0.7	-0.51	-0.66	-0.64	0.28	0.56	0.75
std	0.24	0.24	0.28	0.66	1.4	0.68	0.69	0.86	0.27	0.33	0.37	0.33	1.07	0.65	0.83
min	-1.01	-0.97	-1.18	-0.95	-1.33	-0.88	-1.91	-2.72	-1.13	-1.64	-1.27	-1.12	-0.64	-0.67	-0.86
25%	-0.74	-0.78	-0.9	-0.55	-0.81	0.2	-0.91	-0.62	-0.9	-0.75	-0.92	-0.88	-0.45	0.19	0.06
50%	-0.61	-0.68	-0.76	-0.34	0.87	0.66	-0.48	-0.02	-0.75	-0.53	-0.75	-0.74	-0.13	0.32	0.73
75%	-0.46	-0.56	-0.6	-0.1	1.29	1.13	0.08	0.63	-0.59	-0.27	-0.45	-0.43	0.53	0.96	1.41
max	0.22	0.2	0.22	3.6	4.89	2.65	1.32	1.52	0.29	0.34	0.37	0.29	5.61	2.16	2.98

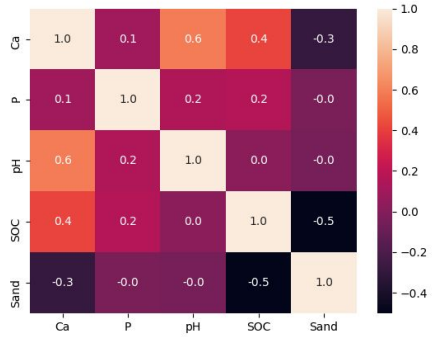
Tabela: Statističke mere prostornih podataka

Atributi o propustljivosti svetlosti, sa druge strane, u proseku nose varijabilitet od 0.15 pa tako postoje observacije koje se u svega nekoliko vrednosti od ukupno 3565 razlikuju od drugih. U narednom delu biće testirano uklanjanje atributa ispod praga varijacije, prebacivanje u prostor glavnih komponenti, ali i regularizovani linearni modeli koji bi znali kako da se nose sa velikim brojem ulaznih atributa.
Ciljni atributi su:

- SOC - količina organskog ugljа u tlu (Soil Organic Carbon),
- pH - vrednost kiselosti zemljišta,
- Ca - količina kalcijuma koja se može ekstrahovati Mehich-3 ekstrakcijom,
- P - količina fosfora koja se može ekstrahovati Mehich-3 ekstrakcijom, i
- Sand - količina peska.

Kako je prikazano na grafiku, između ciljnih atributa postoji određena zavisnost, što ukazuje na to je poželjno probati dva pristupa u učenju višeciljnih modela:

1. Učenje n različitih modela za svaki od n ciljnih atributa;
2. Učenje zajedničkog modela.



Grafik: Korelaciona matrica zavisnosti ciljnih atributa zasnovana Pearson-ovim koeficijentima

2. Preliminarna priprema podataka

Binarni atribut *Depth* prebačen je u numerički mapiranjem vrednosti *Subsoil* u 0, i *Topsoil* u 1. Zasnovano na početnom istraživanju i opisu podataka, postavljene su hipoteze o preprocesiranjima koja bi moguće pomogla izgradnji tačnijeg modela.

- ➔ Hipoteza 1: Uklanjanje spektra m2379.76-m2352.76 neće smanjiti tačnost modela.
- ➔ Hipotezu 2: Skaliranje atributa će imati uticaj na poboljšanje algoritma.
- ➔ Hipoteza 3: Uklanjanje izuzetaka može pomoći izgradnji tačnijeg modela.
- ➔ Hipotezu 4: Neuključivanjem atributa sa malom varijacijom u model može se postići veća tačnost.
- ➔ Hipoteza 5: Učenje modela nad glavnim komponentama datih atributa povećaće se tačnost.
- ➔ Hipoteza 6: Učenje zajedničkog modela višeciljne regresije će dati bolje rezultate predviđanja.

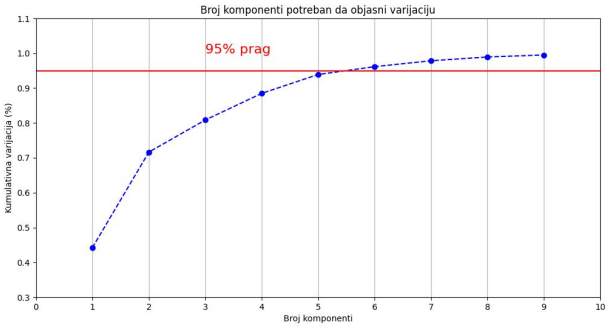
data_index	score_mean
minmax_scaled_spatial_comp+m	0.42
spatial_comp+m	0.44
minmax_scaled_full_wo_co2	0.44
minmax_scaled_full	0.47
full	0.48
full_wo_co2	0.48
minmax_scaled_high_var_x	0.54
full_components	0.60
high_var_x	0.61
minmax_scaled_full_components	0.62
spatial+m_comp	0.62
minmax_scaled_spatial+m_comp	0.67

Tabela: Uprosečena tačnost tri modela za različite pristupe preprocesiranju

Na osnovu rezultata prihvaćene su hipoteze 1, 2, 3, delimično 5 i delimično 6 (kojoj će više pažnje biti posvećeno u nastavku), dok je 4 odbačena.

Na ovaj način utvrđeno je da će se u okviru preprocesiranja, preuzeti sledeći koraci:

1. Uklanjanje redova koji sadrže bar jednu vrednost izuzetka definisanu kao broj koji odstupa više od 5 standardnih devijacija od srednje vrednosti
2. Odvajanje izmerenih od prostornih atributa
3. Uklanjanje izmerenih atributa iz spektra m2379.76-m2352.76
4. Prebacivanje prostornih atributa u prostor glavnih komponenti
5. Kreiranje skupa prediktora kao uniju izmerenih atributa i glavnih komponenti prostornih atributa
6. Skaliranje MinMax normom



Grafik: Broj komponenti potreban da objasni 95% varijacije prostornih atributa

3. Modelovanje

3.1. Korišćene metode evaluacije i glavna metrika

Modeli su evaluirani korišćenjem *sklearn* kros-validacije za adhoc rezultate, dok je za sav detaljan rad korišćena samostalno implementirana kros-validacija. Kao glavna evolutivna metrika korišćena je metrika definisana od strane izvora podataka. U pitanju je srednja vrednost srednje kvadratne grešeke po koloni (eng. *mean columnwise root mean squared error, MCRMSE*):

$$MCRMSE = \frac{1}{5}\sum_{j=1}^{5}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{ij} - \widehat{y}_{ij})^2}$$

3.1. Eksperiment: Inicijalni regresioni modeli

Modeli koji će biti testirani su izabrani tako da bude moguća provera Hipoteze 6, kao i uočene prirode podataka i pretpostavke da je potrebna regularizacija. Izabrani su modeli predstavnici obe grupe modela višeciljne regresije. Za pristup učenju n različitih modela za svaki od n ciljnih atributa kao predstavnici su izabrani Ridge, *Lasso*, *ElasticNet* kao predstavnici linearnih modela, K-nn kao predstavnik modela koji uči o prostoru atributa, ali i modeli obuhvaćeni *MultiOutputRegressor* klasom. Ova *sklearn* implementacija omogućava da se uče različiti modeli za svaki ciljni atribut, tj. ponaša sa kao *wrapper* koji omogućava rad sa višeciljnim problemima. *K-nn* i linearni modeli su podrazumevano implementirani tako da rade kao *MultiOutputRegressor*. Pomoću MOR klase učeni su ansambl modeli *AdaBoost* i *GradientBoostingRegressor*.

Za pristup učenju zajedničkog modela izabrani su predstavnici *DecisionTreeRegressor*, *ExtraTrees* i *RandomForestRegressor*, kao i *RegressorChain K-nn*. Algoritmi koji rade sa stablima poput *DecisionTreeRegressora* vode računa o povezanosti atributa tako što se gradi jedan model za sve n ciljne atribute i optimizuje zajednička funkcija greške tokom učenja. Sa druge strane, klasa koja služi kao *wrapper* za ulančavanje algoritama je *RegressorChain*. Pomoću nje moguće je učiti svaki sledeći model na osnovu ulaznih atributa i izlaza prethodnog modela, tako da i dalje ostaje n modela, ali svi osim prvog sadrže informaciju o drugim ciljnim atributima. Ulančavanje algoritama je implementirano sa bazičnim *K-nnom*.

model	mean_score	std_score
Ridge	**0.22**	**0.03**

K-nn	0.32	0.01
RegChain K-nn	0.32	0.02
Extra trees	0.35	0.07
MultiO/P AdaB	0.4	0.04
RandomForestRegressor	0.42	0.04
MultiO/P GBR	0.56	0.07
Decision Tree Regressor	0.62	0.2
Lasso	0.88	0.15
ElasticNet	0.88	0.15

Tabela: Rezultati inicijalnih modela nad preprocesiranom skupu podataka

Iz tabele se zaključuje da Ridge linearni model najviše odgovara podacima. Sa druge strane, njegovi rezultati češće variraju nego rezultati K-nn, pa zbog je dobro probati optimizovati K-nn kako bi se dobili bolji rezultati. Ostali modeli nisu od velikog značaja što zbog gorih rezultata, što zbog pretreniranosti. Može se takođe zaključiti da se ovaj problem može dobro modelovati i zasebnim modelima poput *Ridge-a* i zajedničkim modelom poput *K-nna*. Upravo ta dva modela su predmet narednog eksperimenta.

3.2. Eksperiment: K-nn model ponderisan Ridge koeficijentima

S obzirom da je Ridge model dao dobre rezultate, dok je K-nn stabilniji, naredna pretpostavka bila je da će K-nn model koji računa distancu ponderisanu Ridge koeficijentima dati rezultate. Model postiže tačnost **0.42**, što je gore od odvojenih modela, tako da je ovaj pristup odbačen.

3.3. Eksperiment: Podešavanje hiperparametra

Za poslednji eksperiment izabrano je da se i dalje zadrže dva model - Ridge kao predstavnik prvog i K-nn kao predstavnik drugog tipa. S obzirom na uočene izazove u postizanju tačnosti evaluacije pri kros-validacije, odlučeno je da se ne koristi implementacija *GridSearcha* iz *sklearna* već da se koristi prilagođen *GridSearch* napisan za potrebe ovog projekta.
Testirana je petostruka kros-validacija sa svim ranije prihvaćenim preprocesiranjima.
Za Ridge algoritam testirano je 20 vrednosti izmedju 0.1 i 2.0. Za K-nn algoritam testiran je broj suseda od 1 do 10 kao i otežavanje instanci po blizini i uniformno. U nastavku su dati rezultati.

alpha	mean_columnwise_root_mse	std
0.1	0.21	0.03
1.8	0.24	0.05
0.5	0.25	0.08
0.7	0.25	0.03

Tabela: Rezultati podešavanja hiper-parametra alpha za Ridge model

weights	k	mean_columnwise_root_mse	std
uniform	9	0.36	0.12
distance	3	0.36	0.08
distance	6	0.36	0.06
uniform	4	0.38	0.11

Tabela: Rezultati podešavanja hiper-parametara težine i broj suseda za K-nn model

Ridge model sa hiper-parametrom alpha=0.1 pokazao se kao najbolji sa postignutim rezultatom ***0.21***, dok je K-nn težio pretreniranju. Odlučeno je da model Ridge bude istreniran na čitavom skupu podataka, sa svim prethodno definisanim preprocesiranjima.

4. Zaključak

Poboljšanje modela pri korišćenju reprezentacije podataka i međuzavisnost ciljnih atributa nagoveštava da je moguće da modeli veštačkih neuronskih mreža mogu postići dobre rezultate. Budući rad bi podrazumevao učenje i prilagođavanje modela neuronskih mreža u službi procenjivanja svojstva zemljišta.

Višeciljni regresioni problemi predstavljaju zanimljiv izazov. Iako standardne korelacione metrike pokazuju zavisnost između ciljnih atributa, zahtevno je ugraditi tu zavisnost u jedan algoritam jer obično sa njom ide i veća pristrasnost. Ovaj rad pokazuje da se izborom jednostavnijih linearnih modela detaljno prilagođenih problemu može postiću konkurentna tačnost algoritma, a zadržati mogućnost lakšeg održavanja i razumevanja zbog jednostavnosti bazičnog modela. Dodatno, ovaj rad naglašava prednost korišćenja nenadgledanih modela za učenje reprezentacije podataka, s tim što je prikazano postizanje veće tačnosti pri učenju modela nad glavnim komponentama grupe ulaznih atributa.

Finalni model naučen je nad svim dostupnim podacima i nad njima postiže $MCRMSE$ od 0.19.