Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 621.39
DOI: https://doi.org/10.24867/02RB02Pejak

KLASIFIKACIJA KULTURA NA SLIKAMA SA SENTINEL-2 SATELITA METODAMA MAŠINSKOG UČENJA

CLASSIFICATION OF CROPS ON SENTINEL-2 IMAGES BY MACHINE LEARNING METHODS

Branislav Pejak, *Fakultet tehničkih nauka, Novi Sad*

Oblast – BIOMEDICINSKO INŽENJERSTVO

Kratak sadržaj – *U okviru rada opisana je klasifikacija poljoprivrednih kultura sa Sentinel-2 satelita uz proširenje skupa obeležja dodatkom dvanaeset vegetacionih indeksa. Podaci predstavljaju vremensku seriju multispektralnih slika područja AP Vojvodine u periodu od marta do septembta za 2016. godinu i preuzeti su sa dve putanje satelita R036 i R136 koje obuhvataju teritoriju AP Vojvodine. Za obuku Random forest (RF) klasifikatora prikupljeni su podaci sa terena o kulturama posejanim na određenom broju parcela. Korišćenje vegetacionih indeksa doprinelo je poboljšanju tačnosti klasifikacije kultura za oko 1 %, čime je u svim eksperimentima ostvarena tačnost od 95 %.*

Ključne reči: *klasifikacija, mašinsko učenje, Sentinel-2, Random forest, vegetacioni indeksi*

Abstract – *The thesis evaluates crops classification methods using Sentinel-2 satellite images with twelve vegetation indexes as additional features. Data includes time-series of multispectral images of AP Vojvodina in the period from March to September 2016, taken from the two corresponding satellite paths, R036 and R136. For the classifier training purposes, the information was collected on the type of crops planted on a number of fields. The vegetation indices contributed to the improvement of the classifier performance of 1 % in all experiments, achieving the overall accuracy of 95 %.*

Keywords: *Classification, Machine Learning, Sentinel-2, Random Forest, Vegetation Indices*

1. UVOD

Klasifikacija kultura [1] zasnovana na podacima sa satelitskih slika predstavlja jedan od najvažnijih izvora informacija o raznvrsnosti kultura koje se uzgajaju u svim poljoprivrednim regionima širom sveta. Ovaj rad predstavlja studiju klasifikacije zasnovane na pikselima koja koristi vremensku seriju multispektralnih slika.

Sam problem karakterišu različiti parametri koji se tiču rezolucije kanala kao i talasnih dužina na kojima senzori satelita očitavaju informacije. Takođe je upotrebljena mogućnost kombinacije određenih kanala radi dobijanja novih informacija u vidu vegetacionih indeksa [2] koji su poslužili za poboljšanje klasifikacije. Klasifikacija poljoprivrednih kultura je sprovedena *Random forest* algoritmom (RF), koji predstavlja ansambalski algoritam zasnovan na stablima odluke. Razmotrena je primena drugačijih načina implementacije ovog algoritma iz razloga što različite kulture nemaju iste faze razvoja u istom vremenskom periodu.

NAPOMENA:
Ovaj rad proistekao je iz master rada čiji mentor je bila dr Tatjana Lončar-Turukalo, vanr. prof.

2. PODACI SA SENTINEL-2 SATELITA

Sentinel-2 [3] je satelit Evropske svemirske agencije koji služi za monitoring Zemljine površine. Sentinel-2 prikuplja podatke pomoću optičkih multispektralnih senzora visoke rezolucije. Ovaj satelit sadrži ukupno 13 spektralnih kanala, od čega su 4 kanala 10-metarske rezolucije, 6 kanala 20-metarske rezolucije i 3 kanala 60-metarske rezolucije. Ovih 13 kanala se nalaze u vidljivom, bliskom infracrvenom i kraktotalasnom infracrvenom delu spektra. Slike koje ovaj satelit pruža su podeljene u granule veličine 100 km^2. Područije AP Vojvodine zahvata 8 granula: 34TCS, 34TCR i 34TCQ iz putanje R136 i 34TDS, 34TDR, 34TDQ, 34TER i 34TEQ iz putanje R036.

3. OPIS BAZE PODATAKA

Da bi se omogućila obuka i verifikacija klasifikatora, odnosno pravljenje modela bilo je neophodno prikupiti podatke sa terena o vrstama kultura i njihovim lokacijama. Prilikom prikupljanja podataka obeleženo je preko 30 različitih kultura. Prvobitni skup podataka činilo je 30 klasa ali pošto nije prikupljen dovoljno veliki skup podataka za sve klase pristupilo se rešenju da se klasifikuje samo 5 najzastupljenijih kultura u Vojvodini, dok su sve ostale kulture grupisane u jednu klasu. Na taj način je formirana baza podataka koja za svaku proizvodnu parcelu, odnosno za svaki piksel sadrži jedinstveni ID broj, geografske koordinate i vrednosti piksela koja predstavljaju očitavanja multispektralnih senzora satelita.

Podaci koji su korišćeni u ovom radu su preuzeti sa dve putanje *Sentinel-2* satelita R036 i R136 koji su snimljeni iznad teritorije Vojvodine u periodu između marta i septembra od kojih je za svaki od piksela izdvojeno po 13 spektralnih kanala. Sa prve putanje R036 iskorišćeno je 8 slika, dok je sa druge R136 putanje iskorišćeno 9 slika.

Veličina vektora obeležja zavisila je od toga da li se za klasifikaciju koristila samo jedna od putanja satelita ili kombinacija obe putanje. Takođe su u svrhu istraživanja sprovedeni eksperimenti sa proširivanjem vektora

obeležja u vidu vegetacionih indeksa koji su formirani kao dodatna obeležja za klasifikaciju. U tabeli 1 predstavljene su sve eksperimentalne postavke.

Tabela 1. *Vektori obeležja koji su upotrebljeni za klasifikaciju*

Putanja	Obeležja	Ukupno obeležja
R036	8 Sentinel slika × 13 kanala + (12 vegetacionih indeksa)	104 + (12)
R136	9 Sentinel slika × 13 kanala + (12 vegetacionih indeksa)	117 + (12)
R036 & R136	8 Sentinel slika × 13 kanala + 9 Sentinel slika × 13 kanala + (12 vegetacionih indeksa)	221 + (12)

4. VEGETACIONI INDEKSI

Vegetacioni indeksi [4] se izračunavaju različitim kombinacijom multispektralnih kanala snimaka sa satelita. U ovom radu postavlja se hipoteza da li korišćenje ovih indeksa kao dodatnih izvora informacija može poboljšati tačnost klasifikacije kultura na osnovu satelitskih snimaka. Vegetacioni indeksi koji su korišćeni kao dodatna obeležja u skupu za obuku su:

- NDVI (engl. *Normalized Difference Vegetation Index*),
- EVI (*engl. Enhanced Vegetation Index*),
- EVI2 (*engl. Enhanced Vegetation Index 2*),
- ARVI (*engl. Atmospherically Resistant Vegetation Index*),
- GAVI (*engl. Green-Adjusted Vegetation Index),*
- SAVI (*engl. Soil-Adjusted Vegetation Index),*
- VDVI (*engl. Visible-Band Difference Vegetation Index),*
- NDMI (*engl. Normalized Difference Moisture Index),*
- NLI (*engl.Non-linear Index),*
- NLI2 (*engl. Non-linear Index 2),*
- MNLI (*engl. Modified Non-linear Index),*
- MNLI2 (*engl. Modified Non-linear Index 2).*

5. ALGORITMI ZA KLASIFIKACIJU

Klasifikacija kultura u ovom radu urađena je pomoću *Random forest* (RF) [5][6] algoritma. RF je metod ansambalskog učenja, gde osnovnu jedinicu u ansamblu čini stablo odluke, a zasnovan je na ideji da više „slabih klasifikatora“ može da se kombinuje i zajedno formira „jak klasifikator“.

Svako stablo odluke unutar ansambla obučava se na *bootstrap* skupu uzoraka. *Bootstrap* označava tehniku formiranja skupa za obuku uzorkovanjem sa vraćanjem iz celokupnog raspoloživog skupa za obuku.

Nakon uzorkovanja svakog elementa, on se vraća u skup i postupak se ponavlja dok ukupan broj uzorkovanih elemenata ne postane jednak broju uzoraka u skupu za obuku celokupnog ansambla. Potrebno je uočiti da se može desiti da jedan trening uzorak bude višestruko ponovljen u *bootstrap* skupu, dok neki trening uzorci mogu u potpunosti da budu izostavljeni.

Svako stablo odluke trenirano na bootstrap skupu će prilikom klasifikacije doneti odluku o labeli uzorka koji se trenutno klasifikuje, dok će konačni izlaz RF klasifikatora biti klasa koja je dobila najveći broj glasova. Ovaj sistem kreiranja ansambla i finalne odluke donešene većinskim glasanjem naziva se *bagging*, što je skraćeno od *bootstrap aggregation.*

Primena RF algoritma realizuje se u dve faze:

- Kreiranje ansambla sačinjenog od stabala odluke
- Korišćenje ansambla za predikciju

5.2.1 Kreiranje ansambla

Postupak kreiranja ansambla:

- Slučajno se izabere *k* obeležja od ukupnog broja obeležja *m*, gde je *k<<m*
- Na osnovu *k* obeležja određuje se čvor *d* pomoću najbolje tačke razdvajanja
- Podeliti čvorove na podčvorove koristeći najbolji presek
- Ponavljati prethodne korake sve dok se ne postigne predefinisana dubina stabla *i*
- Kreirati ansambl ponavljajuću prethodne korake *n* puta kako bi se kreiralo *n* stabala.

5.2.2 Predikcija korišćenjem ansambla

Za predikciju pomoću obučenog klasifikatora koristi se pravilo većinskog glasanja

- Uzima se test uzorak i svako stablo, član ansambla vrši predikciju klase
- Izbroje se glasovi za svaku predviđenu klasu
- Usvaja se klasa koja je dobila najveći broj glasova

Da bi se evaluirale performanse koristeći obučeni RF algoritam, mora se proći kroz sve test uzorke, koristeći pravila odlučivanja svakog kreiranog stabla.

Promenljivi parametri RF algoritma su broj članova ansambla i minimalan broj uzoraka potrebnih za podelu čvora. Ispitivanje tačnosti klasifikacije izvršeno je u zavisnosti od promene broja članova u ansamblu i minimalnog broja uzoraka potrebnih za podelu čvora.

6. OBUKA I EVALUACIJA KLASIFIKATORA NAD EKSPERIMENTALNIM PODACIMA

Da bi se dobili podaci za klasifikaciju bilo je neophodno napraviti binarne maske za svaku sliku sa Sentinel satelita. Na mestima u masci za koje postoje podaci sa terena data je vrednost 1, dok 0 predstavlja neobeležene podatke odnosno mesta na kojima nema informacija o vrsti kulture (klasi). Za ovo istraživanje upotrebljeno je 6 klasa. Kulture od interesa koje su se klasifikovale su: kukuruz, pšenica, soja, šećerna repa i suncokret , dok su u šestu klasu smeštene sve ostale kulture. Takođe su pomoću metode maskiranja eliminisani podaci sa

vrednostima mimo dozvoljenog/podrazumevanog opsega ili podaci koji su imali vrednost piksela 0.

Svaki piksel sadži svoj identifikacioni broj (ID) koji jednoznačno označava pripadnost određenoj parceli. Pikseli koji pripadaju istoj katastarskoj parceli imaju isti identifikacioni broj. ID broj je korišćen i prilikom razvrstavanja uzoraka na trening i test skup.

Primenom tehinike za unakrsnu krosvalidaciju na slučajan način se formira particija sa K-podskupova (u ovom slučaju 10) pri čemu se koristi podatak o klasi kojoj pripada parcela tako da u svakoj particiji bude zastupljeno približno podjednako njiva iz svake klase. Važno je napomenuti da je prilikom krosvalidacije, podela na trening i test uzorke vršena vodeći računa da pikseli sa iste parcele ne budu i u trening i u test skupu. Merenja koja odgovaraju pikselima iz iste parcele su veoma slična pa bi procenjena greška bila manja od stvarne. Zato je prilikom formiranja podskupova za krosvalidaciju slučajna podela urađena na nivou parcela, a ne na nivou samih piksela. Kako parcele nisu istih dimenzija odnosno ne sadrže isti broj piksela ovako formirani poskupovi nisu nužno podjednaki, ali se gledalo da klase budu podjednako zastupljene u obe grupe uzoraka.

Nakon toga vrši se obuka klasifikatora na svim podacima iz trening skupa. Obučen klasifikator se zatim primenjuje na test skup. Pošto se radi o višeklasnom problemu algoritam mašinkog učenja može sprovesti obuku na dva načina. Prvi način je da se algoritam obuči pomoću jednog klasifikatora koji razlikuje sve klase, dok se u drugom slučaju koriste binarni klasifikatori. U slučaju binarne klasifikacije formira se više klasifikatora i konačna odluka se donosi naknadno.

Tehnika koja je razmotrena u ovom radu je *jedan protiv svih* (*engl. One Versus All*, OVA). U tom slučaju se za primarnu klasu proglašava jedna kultura, a sve ostale se smeštaju u sekundarnu. Zatim se identična podela na primarnu i sekundarnu klasu pravi unutar sekundarne klase iz prethodne iteracije i postupak se ponavlja dok u sekundarnom skupu ne ostane samo jedna klasa. Algoritam će tako u ovom slučaju napraviti 6 binarnih klasifikacija za postojećih 6 klasa.

7. REZULTATI

Performanse *Random foret* klasifikatora na Sentinel-2 satelitskim snimcima dobijene nakon krosvalidacije na 10 podskupova date su u tabelama 2, 3, 4 i 5.

Tabela 2. *Tačnost klasifikatora za putanju satelita R036*

		Putanja R036 - tačnost	
Broj članova ansambla	Min procenat uzoraka po čvoru	MS [%]	MS + VI [%]
10	0,1	84,45	87,51
100	0,001	92,57	92,97

Početna ideja je bila da se ispitaju svi mogući slučajevi prilikom pravljenja baze podataka sa podešavanjem parametara za broj uzoraka u svakom čvoru i brojem stabala odlučivanja. U prvom slučaju su korišćeni samo podaci u izvornom obliku dobijeni kao odzivi sa multispektralnih (MS) senzora satelita, dok su u drugom korišćeni kombinovani podaci sa svim vegetacionim indeksima (VI).

Tabela 3. *Tačnost klasifikatora za putanju satelita R136*

		Putanja R136 – tačnost	
Broj članova ansambla	Min procenat uzoraka po čvoru	MS [%]	MS + VI [%]
10	0,1	89,26	89,92
100	0,001	93,51	93,60

Tabela 4. *Tačnost klasifikatora za kombinaciju putanja satelita R036 i R136*

		Putanje R036 i R136 - tačnost	
Broj članova ansambla	Min procenat uzoraka po čvoru	MS [%]	MS + VI [%]
10	0,1	91,02	91,14
100	0,001	94,89	94,93

Ideja za višestruku binarnu klasifikaciju razmotrena je radi prevazilaženja osnovne mane višeklasnog klasifikatora za datu primenu, zbog činjenice da je višeklasni klasifikator moguće pustiti u rad tek nakon završenog vegetacionog ciklusa poslednje kulture. Tako bi se korišćenjem binarnog klasifikatora mapa pšenice mogla dobiti nekoliko meseci ranije. Eksperimenti sa binarnim klasifikatorima izvršeni su sa parametrom „minimalan procenat uzoraka po čvoru“ postavljenim na 0.1 i 0.001 koji je u svim eksperimentima poboljšao klasifikaciju. Urađena je evaluacija sa parametrom „broj članova ansambla“ postavljenim na vrednosti 10 i 100, za obe putanje, kao i za fuziju putanja. Rezultati tačnosti za OVA bianarizaciju prikazani su u tabeli 5.

Tabela 5. *Rezultati tačnosti za* OVA *klasifikaciju primenom Random forest algoritma*

Putanja	Broj članova ansambla	Min procenat uzoraka po čvoru	Tačnost [%]
R036	10	0,1	87,25
	100	0,001	93,02
R136	10	0,1	90,99
	100	0,001	93,91
R036 i R136	10	0,1	90,17
	100	0,001	94,92

Najbolji rezultat ostvaren je na fuziji putanja R036 i R136 i kombinaciji obeležja (multispektralni kanali satelita + vegetacioni indeksi).
Za OVA pristup klasifikaciji razmatrana je tačnost klasifikacije na kombinaciji podataka sa multispekralnih senzora satelita i svim vegetacionim indeksima (MS + VI).

Dobijeni rezultati su pokazali da je selekcija parametara RF algoritma od velikog značaja za poboljšanje performansi samog klasifikatora. Eksperimentalnom pretragom prostora parametara je utvrđeno da najbolje performanse klasifikator ostvaruje ukoliko je broj članova ansambla (broj stabala) postavljen na 100 i ukoliko je minimalan broj uzoraka koji mogu da završe u jednom čvoru 0.001% od ukupnog broja uzoraka.

Uključivanje vegetacionih indeksa u vektor obeležja je poboljšalo rezultat klasifikacije. Tačnost sistema je povećana za oko 1% kako za višeklasni pristup klasifikaciji. Povećanje od 1% nimalo nije zanemarljivo, jer kako se tačnost približava 100% sve ju je teže poboljšati.

Višeklasni i binarni klasifikator su dali veoma slične rezultate prilikom klasifikacije na istim eksperimetima, odnosno uzorcima gde se koriste vrednosti multispektralnih kanala i vrednosti vegetacionih indeksa. Veću tačnost kod pojedinačnih putanja je dala OVA metoda (putanja R036: tačnost = 93.02 %, putanja R136: tačnost = 93.91 %), dok je neznatno veću tačnost imao višeklasni RF klasifikator kod podataka koje predstavljalju kombinaciju obe putanje satelita (R036 i R136: tačnost = 94.93 %). Sa operativnog stanovišta za budući rad značajniji je OVA pristup zasnovan na obuci binarnih klasifikatora za svaku klasu.

9. ZAKLJUČAK

Obrada satelitskih slika useva nalazi primenu u raznim apektima poput nadgledanja zdravlja useva, predikcije prinosa, dobijanja procene za planiranje subvencija, planiranja naredne setvene strukture, izvoza, formiranje cena proizvoda itd. Osnovni zadatak u takvim aplikacijama je da se odredi vrsta useva koja se uzgaja na određenoj parceli.

Kao krajnji rezultat klasifikacija parcela ima značajnu ulogu u pravilnom praćenju i upravljanju obradivim zemljištem kako na lokalnom tako i na globalnom nivou.

Napredak tehnologije omogućio je da ovakve aplikacije budu pristupačnije krajnjim korisnicima usluga, omogućivši značajno povećanje korisnih informacija u pogledu razvoja poljoprivredne proizvodnje. Dobijene informacije mogu biti od velike koristi kako za poljoprivredne proizvođače tako i za industriju kao i za samu državu.

10. LITERATURA

[1] Lugonja P., Marko O., Panić M., Brkljač B., Brdar S., Crnojević V.: Sentinel-2 and Landsat-8 for highresolution land cover mapping in sustainable agriculture, 8. WorldCover Conference, European Space Agency (ESA), GEO, FAO, and EU, Rim: European Space Agency (ESA), 14-16 Mart, 2017

[2] Brkljač B., Lugonja P., Minić V., Brdar S., Crnojević V.: Data enrichment of Sentinel-2 and Landsat-8 surface-reflectance measurements for agriculture oriented services, 3. Earth Observation Open Science Conference, European Space Agency (ESA), Rim: European Space Agency (ESA), 25-28 Septembar, 2017

[3] Sentinel-2 User Handbook, https://sentinel.esa.int/documents/247904/685211/Sentinel-2_User_Handbook

[4] Xue, J. and Su, B., 2017. Significant remote sensing vegetation indices: A review of developments and applications. Journal of Sensors, 2017.

[5] Breiman, L., 2001. Random forests. Machine learning, 45(1), pp.5-32.

[6] James, G., Witten, D., Hastie, T. and Tibshirani, R., 2013. An introduction to statistical learning (Vol. 112). New York: springer.

Kratka biografija:

Branislav Pejak rođen je u Novom Sadu 1994. god. Godine 2017. završava osnovne akademske studije na Departmanu za računarstvo i automatiku, Fakultet tehničkih nauka, Univerziteta u Novom Sadu, smer biomedicinsko inženjerstvo. Iste godine upisuje master akademske studije na istom fakultetu.