Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.9
DOI: https://doi.org/10.24867/29BE13Marinkovic

SEMANTIČKA SEGMENTACIJA POKRIVENOSTI ZEMLJIŠTA U GORNJEM PODUNAVLJU NA SLIKAMA SENTINEL-2 SATELITA

SEMANTIC SEGMENTATION OF LAND COVER IN THE UPPER DANUBE REGION ON SENTINEL-2 SATELLITE IMAGES

Miloš Marinković, *Fakultet tehničkih nauka, Novi Sad*

Oblast – RAČUNARSTVO I AUTMATIKA

Kratak sadržaj – *U okviru ovog rada opisana je semantička segmentacija pokrivenosti zemljišta u Gornjem Podunavlju sa Sentinel-2 satelitskih slika. Segmentacija je izvršena u 6 klasa: krošnje drveća, vodene površine, travnate površine, tlo, poljoprivredno zemljište i vegetacija na vodenim površinama. Podaci predstavljaju multispektralne slike Sentinel-2 satelitskih slika koje obuhvataju područje Gornjeg Podunavlja. Za obuku su korišćeni XG-Boost klasifikator i Random Forest klasifikator. Dronske slike ovog područja služile su za kreiranje činjeničlnog stanja. Kao dodatna obeležja prilikom klasifikacije korišćeni su vegetacioni indeksi. Maksimalna postignuta tačnost iznosila je 93%.*

Ključne reči: *semantička segmentacija, mašinsko učenje, Sentiel-2, Random forest, XG boost, vegetacioni indeksi*

Abstract – *This paper describes the semantic segmentation of land cover in the Upper Danube region using Sentinel-2 satellite images. The segmentation was performed into 6 classes: tree canopies, water bodies, grassland, bare soil, agricultural land, and vegetation on water surfaces. The data utilized multispectral Sentinel-2 satellite images covering the Upper Danube area. XG-Boost and Random Forest classifiers were employed for training. Drone images of the area served to establish ground truth. Vegetation indices were used as additional features during classification. Maximal accuracy was 93%.*

Keywords*: semantic segmentation, machine learning, Sentinel-2, Random Forest, XG Boost, vegetation indices*

1. UVOD

Semantička segmentacija pokrivenosti zemljišta, zasnovana na podacima sa Sentinel-2 satelitskih slika zasniva se na klasifikaciji svakog pojedinačnog piksela nezavisno, upotrebom nekog algoritma mašinskog učenja [1]. Ovaj rad predstavlja primenu ovog pristupa na problem detekcije različitih slojeva koji pokrivaju oblast Gornjeg Podunavlja. Ti slojevi su: krošnje drveća, vodene površine, travnate površine, tlo, poljoprivredno zemljište i vegetacija na vodenim površinama.

Svaki uzorak, u ovom slučaju piksel, opisan sa više kanala različite rezolucije i talasnih dužina na kojima senzori satelita dobijaju informaciju.

NAPOMENA:
Ovaj rad proistekao je iz master rada čiji mentor je bio dr Milan Segedinac, vanredni prof.

Takođe, kao dodatna obeležja korišćeni su vegetacioni indeksi, dobijeni kombinovanjem već postojećih kanala. Vegetacioni indeksi se pokazuju kao vrlo značajni atributi za kvalitetnu klasifikaciju vegetacije [2]. Za klasifikaciju u ovom radu korišćeni su *Random Forest* i *XG Boost* algoritmi.

2. SENTINEL-2 SATELIT i L2A proizvod

Sentinel-2 satelit je deo evropskog programa za praćenje Zemlje, poznatog kao *Copernicus program*, koji pruža visokokvalitetne optičke slike Zemljine površine. Ovaj satelit ima mogućnost snimanja u više spektralnih kanala, što omogućava detaljno praćenje promena na Zemljinoj površini tokom vremena. Osim toga, Sentinel-2 slike su besplatno dostupne za upotrebu u istraživanju, razvoju i komercijalnim aktivnostima.

L2A proizvod u kontekstu daljinskog osmatranja i satelitskih snimaka se odnosi na nivo procesiranja sirovih satelitskih snimaka. U okviru programa Kopernikus, koji uključuje Sentinel satelite, podaci o slikama se obrađuju i distribuiraju u različitim nivoima, pri čemu svaki predstavlja različitu fazu obrade i unapređenja. Proizvod nivoa 2A je srednji nivo obrađenih podataka koji uključuje dodatne korekcije i unapređenja u poređenju sa podacima nivoa 1C.

Podaci nivoa 1C obično sadrže ortorektifikovane slike refleksije na vrhu atmosfere, dok podaci nivoa 2A prolaze kroz dodatnu obradu kako bi pružili refleksije na dnu atmosfere, korekciju atmosfere i dodatna unapređenja poput maskiranja oblaka i senki, korekciju vodene pare i moguću korekciju aerosola.

Proizvod nivoa 2A je posebno koristan za aplikacije koje zahtevaju tačne vrednosti refleksije površine, kao što su praćenje vegetacije, klasifikacija pokrivenosti zemlje i detekcija promena. U ovom radu korišćene su upravo slike sa ovim nivom obrade.

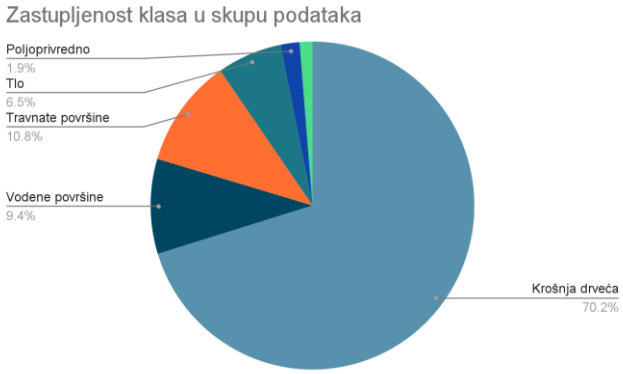
3. Opis baze podataka

Baza podataka predstavlja jednu satelitsku sliku u tri datuma: 2023.06.13, 2023.06.23 i 2023.07.03. Slika ima rezoluciju od 10 metara po pikselu i veličine je 10980x10980, dok je region od interesa veličine 400x400. Svaka slika je mulitspektralna i sadrži 12 kanala različitih delova elektromagnetnog spektra. Pored 12 kanala kao početnih obeležja koriste se i 22 vegetaciona indeksa kao dodatna obeležja.

Sa 34 obeležja u tri datuma dolazi se do ukupnog broja od 102 obeležja koji reprezentuju svaki uzorak odnosno

svaki piksel na slici. Skup podataka nije ujednačen, raspored klasa se može videti na slici 1.

Trening skup je sačinjen od 80% ukupnog uzoračkog skupa, a test skup predstavlja preostalih 20% podataka. Slika koja se obrađuje je slika iz Sentinel-2 putanje R036 sa oznakom T34TCR.



Slika 1. *Zastupljenost klasa u skupu podataka*

4. VEGETACIONI INDEKS

Vegetacioni indeksi su indeksi koji se računaju na osnovu kanala multispektralnih satelitskih slika i nose dodatnu informaciju o tome šta se nalazi na tim slikama.

Neki vegetacioni indeksi imaju različite vrednosti u različitim trenucima u vremenu za različite terene. Ovo su indeksi koji su se koristili:

NDVI *(Normalized Difference Vegetation Index)*

NDVIRE1 *(Normalized Difference Vegetation Index Red Edge 1)*

NDVIRE2 *(Normalized Difference Vegetation Index Red Edge 2)*

NDVIRE3 *(Normalized Difference Vegetation Index Red Edge 3)*

EVI *(Enhanced Vegetation Index)*

EVI2 *(Enhanced Vegetation Index 2)*

VARI *(Visible Atmospherically Resistant Index)*

SAVI *(Soil-Adjusted Vegetation Index)*

ARVI *(Atmospherically Resistant Vegetation Index)*

GAVI *(Green Atmospherically Resistant Vegetation Index)*

VDVI *(Visible Difference Vegetation Index)*

NDWI *(Normalized Difference Water Index)*

NDWI2 *(Normalized Difference Water Index 2)*

NLI *(Normalized Leaf Index)*

NLI2 *(Normalized Leaf Index 2)*

MNLI *(Modified Normalized Leaf Index)*

MNLI2 *(Modified Normalized Leaf Index 2)*

NDMI *(Normalized Difference Moisture Index)*

TG *(Triangular Greenness Index)*

GLI *(Green Leaf Index)*

ExG *(Excess Green Index)*

CIVE *(Color Index of Vegetation Extraction)*

Ovi indeksi su odabrani na osnovu problema koji se rešava i povezanosti između prediktovanih klasa i ovih obeležja.

Slika 2. *Izgled Gornje Podunavlja Sentinel-2 slike*

5. ALGORITMI MAŠINSKOG UČENJA

Ovaj rad predlaže dva pristupa za klasifikaciju pojedinačnih piksela. Kako se suštinski ovde radi o tabelarnim podacima, ideja je da se koriste standardni algoritmi mašinskog učenja kao što su Random Forest i XGBoost algoritam i da se meri njihova makro usrednjen *F1-score*.

5.1. RANDOM FOREST

Random forest [3], koji je predstavio Leo Breiman 2001. godine, predstavlja metodu grupnog učenja koja se koristi za klasifikaciju i regresiju. Funkcioniše tako što tokom obuke gradi više stabala odlučivanja i daje izlaz u vidu moda klasa (za klasifikaciju) ili prosečne vrednosti predikcija (za regresiju) pojedinačnih stabala. Ovaj pristup poboljšava tačnost predikcija i kontroliše prekomerno prilagođavanje (overfitting) prosečavanjem rezultata. Random forests su otporni na šum i mogu da se nose sa velikim skupovima podataka sa visokom dimenzionalnošću, što ih čini popularnim izborom za razne primene u mašinskom učenju.

5.1. XG BOOST

XGBoost *(eXtreme Gradient Boosting)* [4] je algoritam mašinskog učenje koji kombinuje stabla odlučivanja u sklopu metode pojačavanja gradijentom (**engl.** *gradient boosting*). Razvijen je od strane Tianqi Chen-a 2016. godine i popularan je zbog svoje visoke efikasnosti i tačnosti u širokom spektru problema klasifikacije i regresije. XGBoost radi iterativno, dodajući nova stabla u sekvenci, sa svakim novim stablom koje se fokusira na korekciju grešaka prethodnih stabala. Koristi gradijentni spust za minimizaciju gubitka prilikom dodavanja novih stabala, čime se postiže optimizacija performansi i brža konvergencija u odnosu na druge algoritme.

6. OBUKA I EVALUACIJA MODELA MAŠINSKOG UČENJA

Hipoteza koju ovaj radi ispituje jeste da li dodavanjem obeležja iz više datuma doprinosi tačnosti klasifikaciji. Konkretno postavlja se 6 eksperimenata u kom se svaki put model trenira na istom skupu podataka sa smanjenim brojem obeležja izbacivanjem po jednog datuma iz početnog skupa od tri datuma. Ovaj postupak se ponavalja i za Random Forest i za XGBoost. Takođe modeli se testiraju nad istim skupom podataka u svakom od ovih 6 eksperimenata.

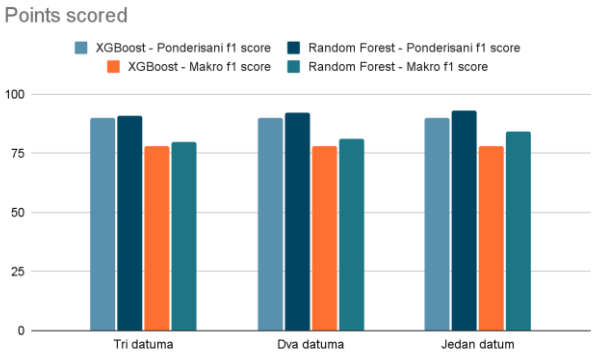
Tabela 1. *Tabela trening i test podataka u svim eksperimentima*

Skup podataka	**Broj piksela**	**E1 - Broj obeležja**	**E2 - Broj obeležja**	**E3 - Broj obeležja**
Trening	**43217**	**34**	**68**	**102**
Test	**10806**	**34**	**68**	**102**

Kako podaci nisu uravnoteženi za evaluaciju modela biće korišćeni ponderisana i makro F1 mera.

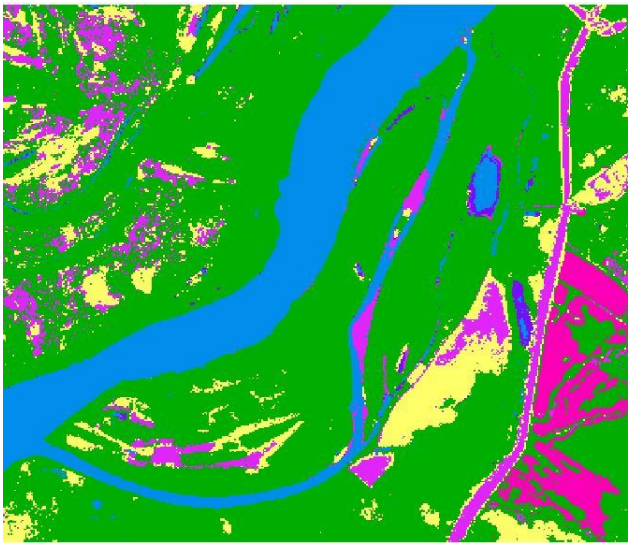
7. REZULTATI

Na osnovu rezultata dobijenih nakon izvršenih eksperimenata hipoteza nije potvrđena, naprotiv pokazuje se da model ima veću tačnost prilikom treniranjem sa obeležjima iz samo jednog datuma. Najbolje se pokazao Random Forest na samo jednom datumu i to je makro F1 mera iznosila 84%, dok je ponderisana F1 mera iznosila 92%. Najlošiji se pokazao XGBoost obučen na sva tri datuma sa 78% makro F1 mere i 90% ponderisane F1 mere.



Slika 3. *F1 mera na osnovu broja obeležja*

Takođe, sa slike se vidi da XGBoost se ne menjaju performanse nezavisno od broja podataka.



Slika 4. *Dobijena mapa predikcija*

8. ZAKLJUČAK

Istraživanje je pokazalo da je Random Forest bez dodatnih obeležje koji se menjaju u vremenu najbolji klasifikator i da vrlo uspešno rešava problem klasifikacije terena i vegetacije sa Sentinel-2 slika. Unapređenja ovog pristupa bi se ogledala u upotrebi dubokih rekurentnih [5] i konvolutivnih neuronskih mreža [6] koje bi imale veći kontekst o okolini piksela kao i povezanosti unutar vremenske serije. Ipak, kako u ovom istraživanju nije postojalno dovoljno obeleženih podataka. takvi pristupi ne bi imali smisla.

9. LITERATURA

[1] Svoboda, J., Štych, P., Laštovička, J., Paluba, D.; Kobliuk, N.: „Random Forest Classification of Land Use, Land-Use Change and Forestry (LULUCF) Using Sentinel-2 Data—A Case Study of Czechia“, Remote Sens. 14, 1189, 2022.
https://doi.org/10.3390/rs14051189

[2] Xue J., Su B.: „Significant Remote Sensing Vegetation Indices: A Review of Developments and Applications“, Journal of Sensors, vol. 2017, Article ID 1353691, 17 pages, 2017.
https://doi.org/10.1155/2017/1353691

[3] Breiman, L.: „Random Forests“, Machine Learning 45:5-32. 2021.
http://dx.doi.org/10.1023/A:1010933404324

[4] Chen, T., Guestrin, C.: „XGBoost: A Scalable Tree Boosting System“, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794, 2016.
https://doi.org/10.1145/2939672.2939785ž

[5] Marhon, S.A., Cameron, C.J.F., Kremer, S.C.: „Recurrent Neural Networks. In: Bianchini, M., Maggini, M., Jain, L. (eds) Handbook on Neural Information Processing“, *Intelligent Systems Reference Library*, vol 49. Springer, Berlin, Heidelberg, 2013.
.https://doi.org/10.1007/978-3-642-36657-4_2

[6] Yamashita, R., Nishio, M., Do, R.K.G. et al.: „Convolutional neural networks: an overview and application in radiology“, *Insights Imaging* 9, 611–629 2018.
https://doi.org/10.1007/s13244-018-0639-9

Kratka biografija:

Miloš Marinković rođen je u Novom Sadu 2000. god. Master rad na Fakultetu tehničkih nauka iz oblasti Elektrotehnike i računarstva – Sematička segmentacija pokrivenosti zemljišta u Gornje Podunavlju na slikama Sentinel-2 satelita odbranio je 2024.god.
kontakt: 1milosmarinkovic1@gmail.com