Klasifikacija gamma čestica generiranih Corsika programom

Ema Dogančić, Anastasija Jezernik, Ana Peterfaj, Maja Tonček

19.4.2019.

Uvodni opis problema

Astronomija je jedna od opservacijskih znanosti u kojoj je čest problem iz velikog broja podataka prepoznati mali broj zanimljivih događaja koji se suptilno razlikuju od pozadinske buke. U ovom projektu ćemo pokušati razraditi problem binarne klasifikacije: na temelju značajki instance signala odrediti je li to gamma signal ili neki signal pozadine.

Koristimo dataset generiran Monte Carlo programom Corsika koji detaljno simulira „zračne tuševe“ (ionizirane čestice i elektromagnetsku radijaciju koje nastaju kada kozmičke zrake prolaze kroz atmosferu). Generirani podaci simuliraju registraciju gamma čestica u Chereknov teleskopu.

Skup podataka se sastoji od 19020 instanci, svaka instanca ima 11 značajki (10 + klasa: 'g' za gamma signal i 'h' za pozadinu). Nema vrijednosti koje nedostaju. Značajke su numeričke (realne vrijednsoti) te predstavljaju parametre generiranih slika (kao što su udaljenost piksela od centra, kut između poluosi elipse i sl.). Detaljnije informacije o podacima se nalaze u eksploratornoj analizi.

Cilj i hipoteze istraživanja problema

U ovom projektu planiramo primijeniti neke od poznatih algoritama klasifikacije na prethodno opisanom skupu podataka, te pomoću mjere uspješnosti klasifikatora odrediti koji je algoritam najbolji.

Pregled dosadašnjih istraživanja

Znanstveni radovi koji su koristili ovaj dataset i na njemu primjenjivali razne metode strojnog učenja su [2], [3] i [4]. U svim trima radovima podaci su podijeljeni približno u omjeru 2:1 na one za treniranje te one za testiranje.

U radu [2] korištene su sljedeće metode:

- klasifikacijska stabla (C5.0, CART i slučajne šume)
- metoda s kernelima
- metoda najbližih susjeda (kNN)
- umjetna neuronska mreža ili ANN (NeuNet package, Neural Network with Switching Units ili NNSU, Group Method Data Handling ili GMDH, Multistart Random Search ili MRS te Mutilayer Perceptron Fit ili MPF)
- composite probabilities
- direct selection
- linearna diskriminantna analiza (LDA)
- metoda potpornih vektora (SVM).

Klasifikacijska stabla, metoda s kernelima i metoda najbližih susjeda su davale slične rezultate. Kod klasifikacijskih stabala bolje rezultate su davale metode s više stabala nego one samo sa jednim, a metoda slučajnih šuma se pokazala boljom od drugih dviju metoda C5.0 i CART, te općenito boljom od skoro svih drugih metoda (u pet od šest parametara kvalitete ostvaruje najbolje rezultate). ANN metode su davale rezultate od ponajboljih do osrednjih, ovisno o odabranoj metodi. Također su se pokazale i osjetljivijima na visoku koreliranost ulaznih parametara.
Metode composite probabilities, direct selection, LDA i SVM su davale lošije rezultate.

U radu [3] se proučava metoda slučajnih šuma ograničenih veličina koja koristi težine (po defaultu ih slučajne šume ne koriste) temeljene na leaf levels of confidence. Rezultati su dani za šume s rangom veličina 20-80 i pokazalo se da ova metoda daje bolje rezultate u odnosu na slučajne šume bez težina, no da se razlika među rezultatima smanjuje kako se veličina šume povećava.

U radu [4] se proučava metoda klasifikacijskih stabala s "mekim" splitovima implementirana u C4.5, ali koja koristi simulirano kaljenje pri treniranju podataka. Metoda je uspoređivana s već spomenutim C5.0 i CART metodama.
Pokazalo se je da ova metoda ima sličnu točnost kao i C5.0 s omekšanim stablima, no da daje puno bolje rezultate u usporedbi s CART metodom s neomekšanim stablima.

Materijali, metodologija i plan istraživanja

Skup ćemo podijeliti na training set i test set u omjeru 80:20 što bi značilo 15216 podataka za učenje i 3804 testnih podataka, no ukoliko rezultati ne budu

zadovoljavajući koristit ćemo "cost sensitive" tehnike poput k-struke unakrsne validacije ili pojedinačne unakrsne validacije. Plan nam je ispitati kakve rezultate daju algoritmi za učenje pod nadzorom - metoda najbližih susjeda (k-nn za nekoliko vrijednosti k-ova), naivni Bayesov klasifikator, slučajne šume, kao i metoda potpornih vektora. Nakon toga ćemo usporediti rezultate međusobno i ako ni jedan model ne bude davao zadovoljavajuće rezultate (80% uspješnosti) implementirat ćemo ansambl (primjerice Bagging) kako bismo poboljšali predikcijsku moć modela.

Podaci su preuzeti sa stranice UCI [1].

Smatra se da je klasificiranje pozadinskog signala kao gamma signala (false positive) gore od klasificiranja gamma signala kao pozadinskog (false negative) [1]. Zbog toga, te zbog relativno balansiranog (65% - 35%) skupa podataka, kao mjeru uspješnosti klasifikatora planiramo koristiti ROC krivulju [5].

Očekivani rezultati predloženog projekta

Kao konačni rezultat projekta predat ćemo rezultate dobivene gore navedenim algoritmima i eventualno nekim dodatnim, usporedbu alogiratama po odabranoj mjeri uspješnosti te zaključak koji je model najbolji za naš primjer. Očekujemo da će slučajne šume dati najbolje rezultate s obzirom da su slične situacije bile i u spomenutim radovima.

Literatura

[1] https://archive.ics.uci.edu/ml/datasets/MAGIC+Gamma+Telescope
[2] https://www.sciencedirect.com/science/article/pii/S0168900203025051
[3] https://www.researchgate.net/publication/303539515_Experimental_study_of_leaf_confidences_for_random_forest
[4] http://uivty.cs.cas.cz/ savicky/papers/softening.pdf
[5] https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/