Izvještaj
Analiza podataka – CIFAR10

Srđan Topić, IN19/2018, topicsrdjan@uns.ac.rs
Nikola Savić, IN35/2018, savic.in35.2018@uns.ac.rs

I. Uvod

Izveštaj se bavi analizom različitih vrsta slika, kao i klasifikacijom gde računar sam različitim klasifikatorima treba da pogodi šta je na slici. Nakon što bi analizirali podatke(slike) i način na koji su te slike predstavljene, moći ćemo da formiramo tri klasifikatora koji bi na osnovu obučavanja na trening podacima mogli da tačno klasifikujemo na test podacima.

II. Baza podataka

CIFAR10 predstavlja bazu podataka koja sadrži 60000 slika svrstanih u 10 kategorija. Kategorije su: avion, automobil, ptica, mačka, jelen, pas, žaba, konj, brod i kamion.

Svaka slika je dimenzije 32x32 piksela i svaki piksel sadrži niz od 3 vrednosti između 0 i 255. Te vrednosti predstavljaju zastupljenost crvene, zelene i plave boje (RGB kanal).

U ovom izveštaju ćemo uporediti tri različita klasifikatora za rešavanje klasifikacionog problema.

III. Klasifikacija

Prilikom učitavanja baze podataka trening skup sadrži 50000 uzoraka dok test skup sadrži 10000 i oba skupa sadrže ravnomerno raspoređene slike tj. trening skup sadrži po 5000 slika iz svake klase dok test skup sadrži po 1000.

Zbog toga što su slike ravnomerno raspodeljene, neće biti potrebe za manipulacijama trening i test skupa da bi se postigla jednaka proporcionalnost slika.

Jedini korak koji se koristio za pripremu podataka je normalizacija vrednosti za RGB kanal tj. deljenjem vrednosti sa 255 smo ih sveli na opseg [0,1].

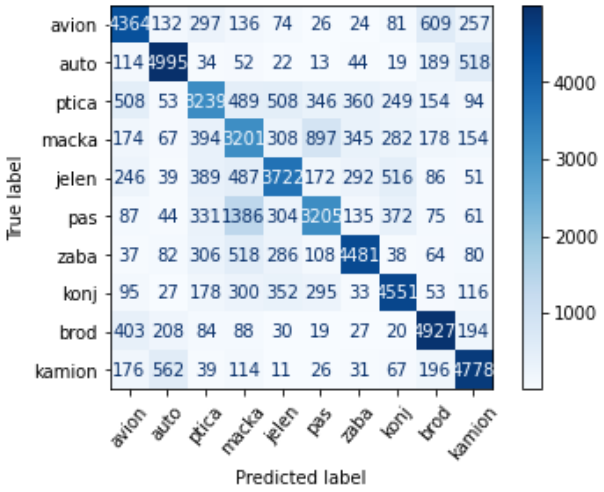
A. Konvoluciona neuralna mreža

Konvoluciona neuralna mreža je klasa veštačke neuralne mreže najčešće korišćena za analizu i klasifikaciju slika. Ovaj tip veštačke neuralne mreže koristi matematičku operaciju konvolucije u jednom od svojih slojeva. Ti slojevi su tako dizajnirani da su pogodni za procesuiranje slika na nivou piksela.

Ova metoda je slična običnoj neuronskoj mreži, samo što pre ulaska u neuronsku mrežu uzorci prođu kroz sloj za izdvajanje osobina. Taj sloj sadrži operacije konvolucije, tj. određivanje filtera za izdvajanje dijelova slike sa kojim se vrši konvolucija, nakon čega se rezultati puštaju kroz ReLu transformaciju, zatim se izvršava Pooling i na kraju se koristi Dropout koji na nasumična mesta postavlja vrednost 0 da bi se izbeglo natprilagođavanje.

Ovim postupkom (slojem za izdvajanje osobina) uspevamo da smanjimo dimenzije slika, izbegnemo natprilagođavanje, da model bude otporniji na manje promene i da ne zavisi od pozicije delova slika(kao npr. glava životinje), a da istovremeno održimo glavne osobine svake slike.

Ovim se rešavaju glavni nedostaci obične neuralne mreže: -previše računanja zbog velikih dimenzija slika
-jednako tretiranje svakog piksela
-osetljivost na poziciju objekata u slici



Sl. 1. Finalna matrica konfuzije primenom unakrsne validacije (cnn)

Unakrsnom validacijom se postiže preciznost oko 70%. Takođe se postiže sličan rezultat korišćenjem klasifikatora na celom skupu.

Preciznost se može povećati promenom parametara kao što su broj epoha, jer što je veći broj epoha to će sigurno preciznost biti veća, ali cena je vremenska zahtevnost. U ovom slučaju je broj epoha 10. Još jedan parametar koji utiče na preciznost je 'dropout', koji na nasumičnim mestima postavlja vrednost 0. Ovaj parametar je uvek poželjno upotrebiti jer se njim izbegava natprilagođavanje.

Na slici 1. u matrici konfuzije vidimo da većini slučajeva model je tačno klasifikovao slike. Doduše postoji

slučaj da u velikim brojevima i pogrešno klasifikuje slike. Npr. klasifikuje sliku kao mačku, a zapravo je pas na slici i obrnuto. Razlog za takve klasifikacije je verovatno kada model vrši izdvajanje osobina slika, pa neke se osobine poklapaju ili su slične kod mačke i psa, kao što su šape, noge i generalno telo. Takođe, razlog može da bude što su slike poprilično male (32x32 piksela), pa je možda teško modelu da izdvoji jasne osobine na slici.

Classification Report:

	precision	recall	f1-score	support
avion	0.76	0.73	0.75	1000
auto	0.86	0.81	0.84	1000
ptica	0.55	0.66	0.60	1000
macka	0.51	0.52	0.52	1000
jelen	0.71	0.60	0.65	1000
pas	0.57	0.67	0.62	1000
zaba	0.81	0.75	0.78	1000
konj	0.80	0.72	0.76	1000
brod	0.81	0.81	0.81	1000
kamion	0.81	0.82	0.82	1000
accuracy			0.71	10000
macro avg	0.72	0.71	0.71	10000
weighted avg	0.72	0.71	0.71	10000

Sl. 2. Izveštaj nakon klasifikacije CNN klasifikatorom

Na slici 2. se može takođe uočiti prethodno komentarisano. Preciznost za pogađanje mačke ili psa je među najnižima u odnosu na ostale klase. To isto smo već zaključili u matrici konfuzije.

B. K najbližih komšija(KNN - K Nearest Neighbours)

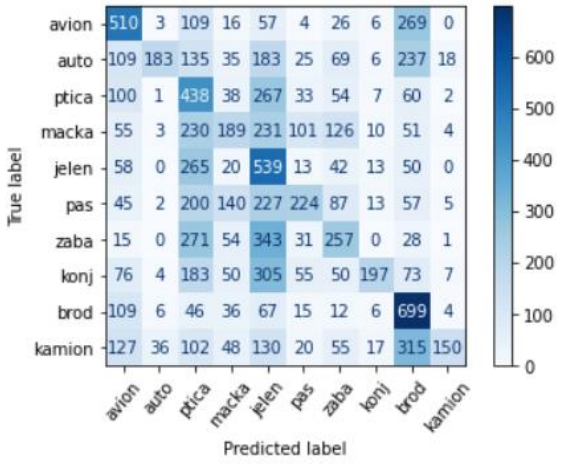
Metoda najbližih komšija je metoda čiji je jedini parametar 'k' tj. koliko najbližih tačaka će model gledati na osnovu čega će vršiti predviđanje.

Testirajući razne metrike pokušali smo da pronađemo onu koja nam je davala nabolje rezultate tačnosti.
Kao rezultate testiranja dobili smo da i euklidska i metrika minskovskog daju jednake rezultate. Tačnost koju daju ove 2 metrike je 46%.

Takođe, osim metrike, testiran je parametar k u knn koji predstavlja parametar koji se odnosi na broj najbližih suseda koje treba uključiti u većinu procesa glasanja. Testirani brojevi za k jesu od broja 1 do 20. Najbolji rezultat se dobija kada se za k uzme broj 10. Isti rezultati se dobijaju i za brojeve od 10 do 20. Tačnost kada se uzme euklidska metrika i za k broj 10 je 46%.

Za ovu metodu je poznato da je nepraktična za ovakvu vrstu klasifikacionog problema. Razlog toga je što KNN gleda svaki pixel slike kao svoje svojstvo i vrednost tj. nema načina da razlikuje neke objekte na slikama i njihove pozicije(kao npr. glava psa). Takođe je nepraktičan kada su velike baze podataka i dimenzionalnosti.

To se može zaključiti na slici 2. gde nam matrica konfuzije pokazuje da je veliki broj uzoraka netačno klasifikovano. Tačnije, kao što smo ranije napomenuli, korišćenjem ovog klasifikatora se dobija preciznost oko 46%, a upotrebom unakrsne validacije, preciznost pada na 34%.



Sl. 3. Matrica konfuzije primenom KNN klasifikatora

Na slici 2. u matrici konfuzije možemo da uočimo da ovaj klasifikator nije pogodan za ovakvu vrstu problema. U velikom broju slučajeva model je vršio pogrešne pretpostavke. Imamo čak i slučajeve gde model klasifikuje pogrešnu klasu više puta nego zapravo pravu.

C. Stablo odluke(Descision tree)

Stablo odlučivanja gradi modele klasifikacije ili regresije u obliku strukture drveta. On rastavlja skup podataka na sve manje i manje podskupove, dok se u isto vreme postepeno razvija povezano stablo odlučivanja. Konačni rezultat je stablo sa čvorovima odluke i čvorovima lista. Čvor odluke ima dve ili više grana. Listni čvor predstavlja klasifikaciju ili odluku. Najviši čvor odluke u stablu koji odgovara najboljem prediktoru koji se zove korenski čvor. Stabla odlučivanja mogu da obrađuju i kategoričke i numeričke podatke.

Analiza stabla klasifikacije je kada je predviđeni ishod klasa (diskretna) kojoj podaci pripadaju.

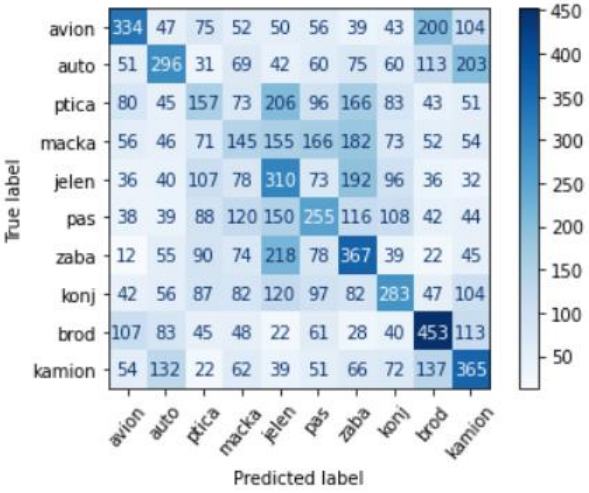
U analizi odluka, stablo odlučivanja se može koristiti za vizuelno i eksplicitno predstavljanje odluka i donošenja odluka. U rudarenju podataka, stablo odlučivanja opisuje podatke (ali rezultujuće stablo klasifikacije može biti ulaz za donošenje odluka).

Testirajući klasifikator stabla odluke na već obradjenim i normalizovanim trening podacima, promenom parametara „max_depth'' dobijaju se različiti rezultati tačnosti. Najveća, ali ne I tako velika tačnost dobija se kada se za ovaj parametar izabere vrednost 10. Ovo smo zaključili kada smo obučavali sa raznim vrednostima ovog parametra, od 5 do 15. Parametar „max_depth'' je u stvari Maksimalna dubina drveta. Ako nije naveden, onda se čvorovi proširuju dok svi listovi ne budu čisti.

Pokušavši da dobijemo bolje rezultate klasifikacije probali smo da izmenimo još neke parametre kao što je npr. „criterion''. Parametar „criterion'' je funkcija za merenje kvaliteta podele. Vrednosti ovog parametra su „gini" za Ginijevu nečistoću i „entropy" za dobijanje informacija.
Korišćenjem vrednosti „gini", koja je I difoltna vresnost

ako se ne navede, dobijaju se bolji rezultati nego korišćenjem vrednosti „entropy“. Tačnost kada koristimo vrednost „entropy“ je 29%.

Klasifikacijom sa klasifikatorom stabla odluke dobijamo tačnost nad trening setom od slabih 31%.



Sl. 4. Matrica konfuzije primenom klasifikatora stabla odluke

Iz ove matrice konfuzije sa slike 4 vidimo da se izdvajaju 3 klase koje ovaj klasifikator malo bolje klasifikuje u odnosu na druge klase. Sliku gde se nalazi žaba klasifikuje 367 puta kao žabu, sliku gde se nalazi brod klasifikuje 453 puta kao brod i sliku gde se nalazi kamion klasifikuje takodje 365 puta kao kamion. Iako ove klase bolje klasifikuje u odnosu na druge klase, vidimo da rezultati i nisu bas najsjaniji.

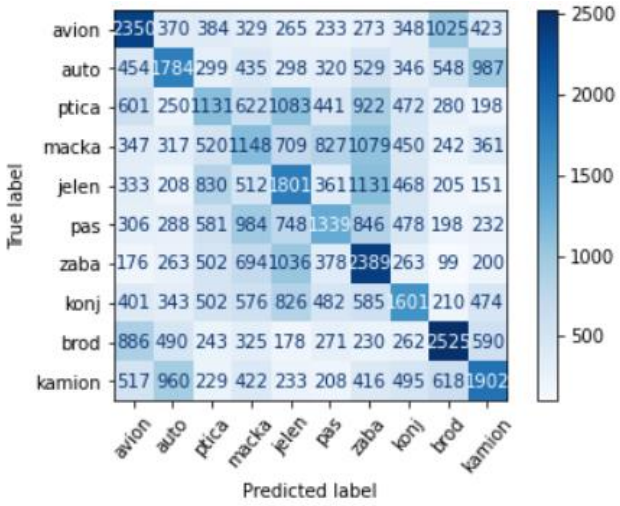
Gledajući sliku 4, vidimo da najveće odstupanje kod klasifikatora stabla odlučivanja ima situacija gde stablo odlučivanja iz klase slika ptice svrstava 206 puta u klasu slika gde se nalazi jelen. Takođe, ne zaostaje ni situacija gde je na slikama automobil, gde ovaj klasifikator klasifikuje pogrešno, takođe kao kamion.

Classification Report:

	precision	recall	f1-score	support
0	0.41	0.33	0.37	1000
1	0.35	0.30	0.32	1000
2	0.20	0.16	0.18	1000
3	0.18	0.14	0.16	1000
4	0.24	0.31	0.27	1000
5	0.26	0.26	0.26	1000
6	0.28	0.37	0.32	1000
7	0.32	0.28	0.30	1000
8	0.40	0.45	0.42	1000
9	0.33	0.36	0.35	1000
accuracy			0.30	10000
macro avg	0.30	0.30	0.29	10000
weighted avg	0.30	0.30	0.29	10000

Sl. 5. Izveštaj nakon klasifikacije stablom odluke

Sa slike 5 zaključujemo da rezultati klasifikacije stablom odlučivanja nisu baš najbolji. Vidimo da je tačnost malih 30%. Najveću preciznost vidimo da je dobio klasifikacijom aviona i broda, gde su preciznosti redom 41% i 40%.



Sl. 6. Matrica konfuzije nakon unakrsne validacije klasifikatorom stabla odluke

U slučaju klasifikacije stablom odlučivanja unakrsnom validacijom vidimo da je procenat pogodjenih uzoraka samo 29.95 %, što je jednako loše u odnosu na procenat pogođenih uzoraka nad test skupom.

Iz ove matrice konfuzije sa slike 6 vidimo da najeći broj pogodjenih uzoraka imamo za klasa avion, žaba i brod, gde redom imamo 2350, 2389 i 2525 tačno pogođenih uzoraka.

U slučaju klasifikatora stabla odlučivanja vidimo da najviše greši u slučaju gde 1036 uzoraka koji pripadaju klasi žaba proglasio da za klasu jelen. Takođe, vidimo da je jelena proglasio kao žabu čak 1131 put i da je veliki broj puta čak 1083 pogrešio kada je za slike gde se nalazi ptica klasifikovao u klase jelena.

D. Upoređivanje klasifikatora i zaključak

Kroz posmatranje svih ovih modela, možemo uočiti da svaki model ima svoje posebne osobine i načine klasifikacije slika.

Jednostavniji modeli, kao što su KNN i stablo odluke, uspevaju neke klase dobro da klasifikuju, ali u velikom broju slučajeva pogrešno klasifikuje.

Najbolji model je definitivno konvoluciona neuralna mreža, čija se glavna dijagonala u matrici konfuzije poprilično izdvaja u odnosu na KNN i stablo odluke.

Možemo doći do zaključka da za ovakvu vrstu problema (klasifikacija slika) je najbolje koristiti konvolucionu neuralnu mrežu koja je superiornija u odnosu na sve ostale modele.