Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.9
DOI: https://doi.org/10.24867/01BE05Rasajski

PRIMENA METODA MAŠINSKOG UČENJA ZA AUTOMATSKU KLASIFIKACIJU MUZIKE PO ŽANRU

AUTOMATIC MUSIC GENRE CLASSIFICATION USING MACHINE LEARNING

Nemanja Rašajski, *Fakultet tehničkih nauka, Novi Sad*

Oblast – ELEKTROTEHNIKA I RAČUNARSTVO

Kratak sadržaj – *Muzički žanrovi su konvencionalne kategorije koje se koriste za opisivanje muzike. Danas se najčešće koriste za klasifikaciju rastućeg broja muzičkih numera, koja bi dalje trebalo da omogući precizniju preporuku i jednostavniju pretragu muzike. U radu je analizirano nekoliko metoda i strategija za automatsku klasifikaciju muzike uključujući konvolucione neuronske mreže (Convolutional neural network – CNN), rekurente neuronske mreže (Reccurent neural network – RNN), mašine potpornih vektora (Support vecotor machines – SVM), random forrest (RF), AdaBoost kao i One vs. Rest (OVR) i klasifikaciju glasanjem. Muzičke numere klasifikovane su na osnovu mel-frequency cepstrum coefficients (MFCC) predstave audio zapisa, a za potrebe CNN-a korišćen je spektrogram. Ostvareni rezultati (~60%) se mogu porediti sa tačnošću (~70%) sa kojom su ljudi u stanju da ispravno procene muzički žanr kao i sa rezultatima ostvarenim u radovima koji su se bavili sličnom temom na istom skupu podataka. Obzirom da preciznost ostvarena u radu nije daleko od procene ljudi, metode bi mogle naći primenu u automatskoj klasifikaciji muzike za potrebe radio stanica ili web sajtova koji se bave distribuiranjem i preporukom muzičkih numera.*

Ključne reči: *klasifikacija muzike po žanrovima, mašinsko učenje, GTZAN skup podataka.*

Abstract – *Music genres are conventional categories that are used for describing music. Today they are most often used for classifying growing music collections, for easier access and recommendation. This paper has analyzed a number of methods for automatic music classification, which include convolutional neural networks (CNN), recurrent neural networks (RNN), support vector machines (SVM), random forests (RF), AdaBoost, voting classifier and one versus rest (OVR). Features used for audio classification were created using mel-frequency cepstrum coefficients (MFCC), and spectrograms which were used in combination with CNNs. The accuracy achieved (~60%) was not at the human level (~70%), but it was not too far off, and it is in line with other similar approaches. Thus, methods presented in this paper can be used for automatic classification of music, by radio stations or web portals that distribute and recommend music.*

Keywords: *Automatic music genre classification, machine learning, GTZAN dataset*

NAPOMENA:
Ovaj rad proistekao je iz master rada čiji mentor je bio dr Aleksandar Kovačević, vanr. prof.

1. UVOD

Razvoj interneta i prenosnih medija omogućio je korisnicima pristup velikim količinama multimedijalnog sadržaja. Da bi se organizovale i pretražile rastuće muzičke kolekcije, potrebni su automatizovani alati za filtriranje i procesiranje audio fajlova. Neke od ovih funkcionalnosti mogu biti olakšane pomoću metapodataka, koji pružaju dodatne informacije o sadržaju. Često ti metapodaci nisu dostupni, pa je potrebno analizirati podatke bez njih. Tada se neophodne informacije o pesmama uzimaju direktno iz audio fajlova. Takve informacije mogle bi obuhvatati žanr, raspoloženje, stil, izvođača. U ovom radu akcenat je na indentifikovanje žanra pesme iz grupe od 10. Većina sistema za klasifikaciju muzike imaju dva koraka, izvlačenje osobina na osnovu kojih se vrši klasifikacija i sama klasifikacija. Za svrhe klasifikacije koriste su razne osobine signala, poput zero-crossing, bandwidth, spetral centroid. Skup osobina korišćen za klasifikaciju u ovom radu je MFCC [2], a korišćeni su i spektrogrami (spectrogram) [4] koji su grafička predstava spektruma kao i hromagram [6] (chromagram).

Isprobane su metode predložene u prethodnim radovima koji su se bavili ovom temom, poput SVM (Support vector machines) [8]. Takođe su istražene varijacije gore napomenutih metoda, poput SVM ansambla, kao i u ovoj oblasti slabije istraženi algoritmi poput RNN, CNN.

Skup podataka korišćen za izradu rada je GTZAN Dataset – kolekcija 1000 audio fajlova.

U poglavlju II, biće dat pregled postojećih rešenja. Poglavlje III detaljnije objašnjava skup podataka, kao i sve transformacije nad njim vršene za potrebe ovog rada. U poglavlju IV će biti reči o metodama korišćenim u procesu klasifikacije. Poglavlje V daje pregled i diskusiju rezultata. Poglavlje VI nudi zaključke koje je autor stekao.

2. PREGLED POSTOJEĆE RELEVANTNE LITERATURE

Algoritmi za automatsko prepoznavanje muzičkog žanra opisani su u [3]. U radu je takođe predložen skup svojstava za predstavu muzike, a performanse skupa svojstava evaluirane su treniranjem klasifikatora baziranih na statističkom prepoznavanju šablona. Za klasifikaciju je kreiran GTZAN skup podataka koji se koristi i u ovom radu i dobijena je tačnost od 61% korišćenjem Expectation-maximization (EM) i k-nearest neighbours (KNN) algoritma.

Analiza GTZAN skupa podataka izložena je u [5]. U radu je navedeno da 7.2% audio fajlova potiču iz istog

muzičkog zapisa, uključujući 5% audio fajlova koji su potpuno identični, te mogu dovesti do pogrešne predstave rezultata (npr. ukoliko se isti zapis koristi i za treniranje i za testiranje). Dalje, autori tvrde da je 10.6% audio fajlova pogrešno označeno vodeći se načinom na koji su pesme i autori označeni na sajtu last.fm, kao i da pojedini izvođači dominiraju određenim žanrom (npr. 34% audio fajlova iz rege žanra izvodi Bob Marley) pa se postavlja pitanje iskoristivosti treniranih klasifikatora za klasifikaciju muzike ostalih izvođača.

Radovi predstavljaju dobru osnovu za dalji rad i daju uvid u to kakve rezultate treba očekivati, kao i koja su ograničenja GTZAN skupa podataka.

3. SKUP PODATAKA

Za izradu rada korišćen je GTZAN skup podataka [1] koji sadrži 1000 audio fajlova gde je trajanje svakog fajla 30 sekundi. Skup podataka je podeljen na 10 kategorija (bluz, klasična, metal, kantri, džez, pop, rok, rege, hiphop i dens muzika), a za svaku kategoriju postoji 100 muzičkih numera. Sve numere su 22050 Hz monofoni 16-bit audio fajlovi u AU formatu. Podaci su konvertovani u WAV format. Na osnovu liste date u [5] uklonjeni su svi istovetni audio fajlovi iz skupa podataka, a klase izbalansirane nasumičnim izostavljanjem primera. Tako modifikovani skup podataka sadrži 850 audio fajlova - po 85 audio fajlova za svaki od žanrova, odnosno nešto više od 7 sati ukupnog materijala.

Za treniranje klasifikatora korišćene su tri reprezentacije audio signala - MFCC, hromagram (chromagram) i spektrogram (spectrogram).

3.1 MFCC

Mel-frequency cepstrum (MFCC) reprezentacija je bazirana na načinu na koji ljudi doživljavaju muziku. MFCC sadrži dva tipa filtera koji su raspoređeni linearno na niskim frekvencijama ispod 1000 Hz i logaritamski na frekvencijama iznad 1000 Hz.

Proces pretvaranja audio signala u MFCC reprezentaciju obuhvata šest koraka [2]:

1. *Isticanje (Pre-emphasis)*

Obuhvata proces prosleđivanja signala kroz filter koji naglašava visoke frekvencije

2. *Kadriranje (Framing)*

Signal se deli na frejmove od N uzoraka. Susedni frejmovi se dele na M podfrejmova gde je M < N. Tipične vrednosti koje se koriste su M=100 i N=256.

3. *Okno Heminga (Hamming window)*

Jednačina glasi:

$$Y(n) = X(n) \times W(n) \qquad (1)$$

gde je Y(n) izlazni signal, X(n) ulazni signal, a W(n) se definiše kao:

$$W(n) = 0.54 - 0.46cos\left(\frac{2\pi n}{N} - 1\right)\ 0 \le n \le N - 1 \qquad (2)$$

gde je N širina izražena u semplovima.

4. *Brza Furijeova transformacija (Fast Fourier Transformation)*

Svaki frejm od N semplova se konvertuje iz vremenskog domena u frekventni domen.

5. *Procesiranje skupa mel filtera (Mel Filter Bank processing)*

Faza obuhvata izračunavanje težinske sume spektralnih komponenti filtera tako da se izlaz može aproksimirati Mel skalom. Sledeća jednačina se koristi za izračunavanje opsega (mel) za svaku frekvenciju f u Hz:

$$F(Mel) = [2595 * log10[1 + f]700] \qquad (3)$$

6. *Diskretna kosinusna transformacija*

Proces podrazumeva konvertovanje logaritamskog Mel spektra u vremenski domen korišćenjem diskretne kosinusne transformacije.Rezultat konverovatnja je MFCC. Skup koeficijenata predstavlja akustični vektor.

3.2 HROMAGRAM

Hroma (chroma) je jedna od dve komponente koje se mogu dobiti iz visine zvuka. U poređenju sa tonovima u muzici, hroma predstavlja elemente u skupu {C, $C^{\#}$, D,...,B}, gde oznake $C^{\#}$ i D^{b} odgovaraju istoj hromi. Klasa visine zvuka je definisana kao skup svih tonova koji dele istu hromu.

Reprezentacija hrome se dobija tako što se svakoj klasi dodeljuju tonovi kojima oni pripadaju. Tačnije, audio signal se prvo rastavlja na 88 podgrupa koristeci filtere, izračunava se STMSP (Short-time mean-square power) za svaku podgrupu, i potom se sabiraju svi STMPS koji pripadaju istoj klasi. Ovo rezultuje u realnom 12-dimenzionalnom vektoru, gde v(1) odgovara hromi C, v(2) hromi $C^{\#}$ i tako dalje. Primena ove metode na različitim vremenskim intervalima se naziva hromagram [6].

3.3 SPEKTROGRAM

Spektorgram je grafička reprezentacija spektra frekvencija u zvučnom ili drugom signalu, po vremenu ili nekoj drugoj promenljivoj. Često se koriste u oblasti analiziranja muzike.

Spektrogrami se predstavljaju u formi grafika, gde horizontalna osa predstavlja vreme, a vertikalna osa frekvenciju, treća dimenzija koja predstavlja amplitudu određene frekvencije u određeno vreme, predstavljena je intenzitetom boje na svakoj tački slike.

Spektrogrami se obično kreiraju na dva načina. Aproksimacija pomoću filterbanka (niz band-pass filtera), ili pomoću Fourierove transformacije [7]. U ovom radu spektrogrami su generisani pomoću Fourierove transoformacije. Proces kreiranja spektrograma odgovara računanju, kvadrat apsolutne vrednosti STFT (Short-time Fourier transform) signala *s(t)* za interval *ω* [4]*:*

$$\text{spectrogram}(t,\omega) = |\text{STFT}(t,\omega)|^2 \qquad (4)$$

4. METOD

S ciljem klasifikacije muzike korišćeno je nekoliko algoritama i strategija. Evaluacija algoritama nad istim skupom podataka pružila je uvid u iskoristivost datih algoritama za rešavanje problema klasifikacije muzike. Razlozi za njihovo dati su u nastavku.

4.1 CNN

Konvolutivne neuronske mreže su našle veliku primenu u oblasti klasifikacije slika, gde daju izuzetne rezultate (iznad 95% na ImageNet takmičenju, koje zahteva

klasifikaciju na 200 klasa). Tema ovog rada je klasifikacija muzike na osnovu audio signala, gledajući da se i on može grafički predstaviti pomoću spektrograma. CNN je izabrana u pokušaju da se problem klasifikacije pesama, pretvori u problem klasifikacije slika i primeni algoritam koji se pokazao kao veoma uspešan u tom polju.

4.2 RNN

Pretpostavka je da bi rekurentna neuronska mreža bila u stanju da izmodeluje temporalne zavisnosti među semplovima.

Za treniranje rekurentne neuronske mreže korišćena je NEAT metoda. Metod se zasniva na evoluciji neuronske mreže kroz promenu kako topologije mreže tako i težina. [9]. Kroz proces koji obuhvata enkodiranje, mutaciju, ukrštanje, selekciju, praćenje najboljih gena i očuvanje inovativnih rešenja, algoritam je u stanju da od inicijalne neuronske mreže kreira mrežu koja je u stanju da reši problem za koji je evoluirana.

4.3 AdaBoost

Dve karakteristike AdaBoost algoritma su otpornost na overfitovanje i osetljivost na šumove.. Sa jedne strane, otpornost na overfitovanje će omogućiti bolju generalizaciju što je esencijalno za uspšeno klasifikovanje muzike, dok sa druge strane, osetljivost na šumove može predstavljati prepreku (što će se kasnije i primetiti u rezultatima) sa obzirom na raznolikost u numerama. Za slabe klasifikatore odabrani su stabla odlučivanja zato što se smatra da oni daju najbolje rezultate.

4.4 Random Forest

Razlog za korišćenje Random forest algoritma leži u njegovoj robustnosti na autlajere i šum, brzini izvršavanja obzirom da ga je jednostavno paralelizovati i tačnosti koja je jednako dobra ili bolja od Adaboost algoritma [10].

4.5 SVM

SVM je algoritam koji se najčešće spominje u literaturi kada se rešava problem klasifikovanja muzike, rezultati dobijeni njegovim korišćenjem su među najboljim i zbog toga je procenjeno da je potrebno i njega primeniti.

4.6 One vs. Rest

One vs. Rest strategija korišćena je s ciljem postizanja boljih rezultata kod Random Forest, SVM i AdaBoost klasifikatora.

Metoda obuhvata formiranje deset birnarnih klasifikatora, balansiranje klasa, treniranje svih klasifikatora, njihovo evaluiranje i preuzimanje rešenja onog klasifikatora sa najvećom tačnošću.

Formianje klasifikatora podrazumeva kreiranje klasifikatora za svaki od žanrova (npr. klasifikator koji klasifikuje podatke na džez i nije džez ili metal i nije metal). Za treniranje je korišćeno 136 fajlova, a za testiranje 34 fajlova gde je odnos broja pozitivnih i negativnih primera 1:1 kako u skupu podataka za treniranje tako i u skupu podataka za testiranje.

Sva tri klasifikatora (RF, AdaBoost, SVM) su trenirana i testirana nad istim podacima.

4.7 Većinsko glasanje (Voting classifier)

Razlog za korišćenje većinskog glasanja leži u kombinovanju konceptualno različih klasifikatora za predviđanje klasa glasanjem. Kada se radi o klasifikatorima koji rezultuju sličnim performansama, većinsko glasanje može biti dobar izbor jer omogućava balansiranje nedostataka ovih klasifikatora. U ovom slučaju, za većinsko glasanje korišćeni su AdaBoost, SVM i Random Forest klasifikatori jer su pojedinačno dali slične rezultate.

5. REZULTATI I DISKUSIJA

Kako je istraživanje [11] pokazalo da su ljudi u stanju da sa 70% tačnosti procene žanr muzike, težilo se pronalaženju metode koja bi dostigla približnu tačnost ne bi li se takva metoda mogla koristiti za automatsku klasifikaciju muzike.

Sve metode testirane su na modifikovanom GTZAN skupu podataka (duplikati su uklonjeni, a klase izbalansirane nakon uklanjanja) korišćenjem 5-struke unakrsne validacije (5-fold cross validation). Svi audio fajlovi su iz AU formata konvertovani u WAV format. Nakon toga, izvršeno je konvertovanje WAV fajlova u CSV fajlove koji sadrže MFCC predstavu audio signala odnosno hromagram. Eksperimenti su pokazali da je MFCC dao dvostruko bolje rezultate od hromagrama pa su rezultati isloženi u nastavku nastali klasifikacijom audio fajlova na osnovu MFCC (s izuzetkom CNN-a gde je korišćen spektrogram). Za treniranje i testiranje korišćeno je prvih 12900 semplova što odgovara maksimalnom broju semplova pojedinih fajlova. Izuzetak je rekurentna neuronska mreža koja je zbog hardverskih ograničenja morala biti trenirana korišćenjem 2500 semplova obzirom da je trebalo voditi računa i o sekvencama.

Kao mera evaluacije korišćena je F mera, konkretno implementacija F mere iz biblioteke sklearn [12], micro i macro varijacije. U Tabeli 1 prikazani su rezultati svih testiranih metoda

Kao što se može videti, svi klasifikatori su dali približno iste rezultate, s izuzetkom RNN-a gde je skup podataka morao biti pet puta smanjen zbog hardverskih ograničenja. One vs. Rest strategija doprinela je povećanju tačnosti SVM metode, ali je negativno uticala na AdaBoost metodu. Razlog za to mogu je lošija tačnost binarnih AdaBoost klasifikatora u odnosu na Random Forest i SVM. Na samom kraju, glasanje klasifikatora dalo je najveću tačnost od 62% što odgovara rezultatu ostvarenom u [3] (61%) korišćenjem EM i KNN metode.
Ono što treba imati u vidu je da za potrebe treniranja i testiranja klasifikatora nije korišćen ceo skup podataka već 85% njega. Razlog za to su ponovljeni audio fajlovi - problem GTZAN skupa podataka [5]. U istom radu navedena su i neslaganja sa načinom na koji su audio fajlovi klasifikovani, ali je taj iskaz zanemaren vodeći se mišljenjem da razlog za pripisivanje muzičnje numere žanru može biti subjektivan.

Radi ispitivanja uticaj trajanja trening skupa na rezultate, klasifikatori su trenirani koristeći audio fajlove trajanja od 15s. Sa dvostruko manje semplova, rezultati su bili 47%

lošiji. Takođe, korišćenje hromagrama umesto MFCC-a rezultovalo je tačnošću klasifikatora koja se kretala oko 30%. Iz gore predstavljenog može se zaključiti da bi drugačija predstava audio signala kao i veći broj numera ili duže trajanje numere pozitivno uticalo na tačnost klasifikatora.

Tabela 1 – *Rezultati klasifikacije*

	macro F-score	micro F-score
CNN	0.58	0.59
RNN	0.21	0.23
Random Forest	0.57	0.59
AdaBoost	0.54	0.56
SVM	0.50	0.53
One vs. Rest		
SVM	0.55	0.59
Random Forest	0.51	0.60
AdaBoost	0.32	0.35
Voting Classifier		
SVM+RF+AB	0.57	0.62

6. ZAKLJUČAK

Tema ovog rada je bio problem automatske klasifikacije muzike, radi olakšavanja obrade i labeliranja rastućih muzičkih kolekcija. U radu je korišćeno nekoliko različitih klasifikatora, kao i kombinacije klasifikatora, koji su obučavani korišćenjem GTZAN skupa podataka. Prilikom obučavanja pokazalo se da ovaj skup podataka ima nekoliko nedostataka. Osobine audio fajlova korišćene za klasifikaciju su izvedene pomoću MFCC-a i spektrograma. Svi uspešno obučeni klasifikatori su imali približno istu tačnost od oko 60%.

Iako klasifikatori nisu postigli ljudsku tačnost, razlika između ljudi i klasifikatora nije toliko velika i postoji mogućnost da se može nadoknaditi korišćenjem deskriptivnijeg skupa osobina, većih skupova podataka i kompleksniji arhitektura klasifikatora.

Dalji pravci razvoja bi se prvenstveno odnosili predstavu signala, odnosno izvođenje osobina za klasifikaciju iz signala, gde bi se mogle koristiti i neke primitivnije karakteristike poput, boje zvuka (timbre), ritmičkog sadržaja, visine tona (pitch), samostalno ili u kombinaciji sa MFCC i sličnim skupovima koeficijenata. Druga mogućnost je veći i potencijalno preciznije labeliran (mišljenje većine) skup podataka. Na kraju mogle bi se istražiti i upotrebiti složenije arhitekture trenutno korišćenih klasifikatora, kao i njihove kombinacije.

7. LITERATURA

[1] http://marsyasweb.appspot.com/download/data_sets/, Datum pristupa: 25.3.2017.

[2] Muda, Lindasalwa, Mumtaj Begam, and Irraivan Elamvazuthi. "Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques." arXiv preprint arXiv:1003.4083 (2010).

[3] Tzanetakis, George, and Perry Cook. "Musical genre classification of audio signals." IEEE Transactions on speech and audio processing 10.5 (2002): 293-302.

[4] http://zone.ni.com/reference/en-XX/help/371361E-01/lvanls/stft_spectrogram_core/#details, Datum pristupa: 15.8.2018.

[5] Sturm, Bob L. "An analysis of the GTZAN music genre dataset." Proceedings of the second international ACM workshop on Music information retrieval with user-centered and multimodal strategies. ACM, 2012.

[6] Müller, Meinard. *Information retrieval for music and motion*. Vol. 2. Heidelberg: Springer, 2007.

[7] https://ccrma.stanford.edu/~jos/sasp/, Datum pristupa: 15.8.2018.

[8] Mandel, Michael I., and Dan Ellis. "Song-Level Features and Support Vector Machines for Music Classification." ISMIR. Vol. 2005. 2005.

[9] Stanley, Kenneth O., and Risto Miikkulainen. "Evolving neural networks through augmenting topologies." Evolutionary computation 10.2 (2002): 99-127.

[10] Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.

[11] D. Perrot and R. Gjerdigen, “Scanning the dial: An exploration of factors in identification of musical style,” in Proc. Soc. Music Perception Cognition, 1999

[12] http://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html#sklearn.metrics.f1_score, Datum pristupa: 15.8.2018.

8. BIOGRAFIJA

Nemanja Rašajski rođen je 1993. godine u Novom Sadu. Osnovnu školu ,,Vasa Stajić“ u Novom Sadu završio je 2008. godine, Gimnaziju ,,Isidora Sekulić“ u Novom Sadu, prirodno-matematički smer završio je 2012. godine. Iste godine upisao je osnovne akademske studije na smeru Računarstvo i automatiku, Fakulteta tehničkih nauka u Novom Sadu. Zvanje diplomirani inženjer elektrotehnike i računarstva stekao je 2016. godine, sa prosečnom ocenom 9.70. Iste godine upisao je master akademske studije na smeru Softversko inženjerstvo i informacione tehnologije Fakulteta tehničkih nauka u Novom Sadu. Uža specijalnost na master studijama bila je inteligntni sistemi.