Sprawozdanie - projekt indywidualny

Weronika Zbierowska

Marzec 2023

1 Cel projektu

Projekt *"Rozpoznawanie gatunków muzycznych z wykorzystaniem technik uczenia maszynowego"* realizowany jest w ramach przedmiotu Projekt indywidualny na 4. semestrze studiów inżynierskich Informatyka stosowana na Wydziale Elektrycznym Politechniki Warszawskiej.

Celem projektu jest stworzenie aplikacji dokonującej klasyfikacji gatunków utworów muzycznych. W ramach projektu należy stworzyć własną bazę utworów, dokonać ekstrakcji cech oraz na ich podstawie wytrenować i wdrożyć model uczenia maszynowego.

2 Ekstrakcja cech

Pierwszy etap projektu polegał na zapoznaniu się z pakietem `librosa` oraz utworzeniu prototypu ramki danych zawierającego cechy wyekstrahowane z 3 przykładowych utworów muzycznych.

2.1 Brane pod uwagę cechy

Każdy utwór został zapisany jako zbiór próbek wartości sygnału w danych chwilach czasowych. Aby można było poddać je analizie, dokonano ekstrakcji cech dla każdej ramki czasowej (oprócz tempa). Wynikiem były wektory cech długości liczby ramek. Następnie dla każdego wektora wyliczono średnią oraz odchylenie standardowe. Dla cech, których wartości były analizowane dla kilku pasm częstotliwości (kilka wektorów na 1 cechę) wyliczono średnią i odchylenie standardowe oddzielnie dla każdego pasma.
Stosowana notacja:

- x_r - 1 ramka czasowa
- N - długość 1 ramki
- $n \in \langle 0, N \rangle$ - chwila czasowa w ramce
- $x_r(n)$ - wartość sygnału dla chwili n w ramce x_r
- $X_r(k)$ - wartość amplitudy w chwili k dla częstotliwości $f(k)$

2.1.1 Tempo

Estymowana szybkość utworu wyrażona w uderzeniach na minutę.

2.1.2 Energia krótkoczasowa (ang. Short-time enegry)

Miara określająca energię sygnału w krótkim przedziale czasowym. Dla sygnałów dźwiękowych określa głośność lub intensywność sygnału.

$$STE_r = \frac{1}{N} \sum_{n=1}^{N} |x_r(n)|^2 \tag{1}$$

2.1.3 Wartość skuteczna energii (ang. Root mean square energy)

Miara określająca wartość skuteczną energię sygnału w krótkim przedziale czasowym. Jest równa pierwiastkowi z energii krótkoczasowej.

$$RMSE_r = \sqrt{\frac{1}{N} \sum_{n=1}^{N} |x_r(n)|^2} \tag{2}$$

2.1.4 Współczynnik przekroczeń zera (ang. Zero-crossing rate)

Współczynnik określający liczbę zmiany znaku sygnału na przestrzeni ramki czasowej. Jest silnie skorelowany ze średnią częstotliwością dużej koncentracji energii.

$$RMSE_r = \frac{1}{2} \sum_{n=1}^{N} |sign(x_r(n)) - sign(x_r - 1(n))|, \tag{3}$$

gdzie:

$$sign(x) = \begin{cases} -1, \text{ x} \geq 0 \\ 1, \text{ x} \leq 0 \end{cases} \tag{4}$$

W pakiecie `librosa` współczynnik przekroczeń zera jest zdefiniowany jako ułamek liczby zmian znaku sygnału na przestrzeni ramki do długości ramki, czyli:

$$RMSE_r = \frac{1}{2N} \sum_{n=1}^{N} |sign(x_r(n)) - sign(x_r - 1(n))| \tag{5}$$

Tej więc definicji będę używać.

2.1.5 Centroid widmowy (ang. Spectral centroid)

Środek ciężkości widma sygnału, czyli punkt, w którym koncentruje się większość energii widma. Jest silnie skorelowany z odbieraną przez człowieka jasnością dźwięku.

$$C_r = \frac{\sum_{k=1}^{\frac{N}{2}} f(k)|X_r(k)|}{\sum_{k=1}^{\frac{N}{2}} |X_r(k)|} \tag{6}$$

2.1.6 Szerokość pasma widma (ang. Spectral bandwidth)

Miara rozpiętości widmowej na przestrzeni ramki czasowej.

$$B_r = \left(\sum_{k=1}^{\frac{N}{2}} |X_r(k)|(f(k) - C_r)^p \right)^{\frac{1}{p}} \tag{7}$$

2.1.7 Opadanie widma (ang. Spectral roll-off)

Częstotliwość na przestrzeni ramki czasowej, poniżej której leży 85% całkowitej energii widma. Określa jak szybko energia widma maleje wraz ze wzrostem częstotliwości.

2.1.8 Kontrast widmowy (ang. Spectral contrast)

Kontrast energii dla 6 pasm częstotliwości na przestrzeni ramki czasowej. Estymowany poprzez porównanie średniej energii szczytowej oraz energii w dolnym kwantylu. Im wyższa wartość kontrastu widmowego, tym mniej szumu w sygnale.

2.1.9 MFCC (ang. Mel-frequency cepstral coefficients)

Współczynniki reprezentujące intensywność występowania sygnału w 20 różnych pasmach częstotliwości. Są one wyrażone w skali mel, która jest dostosowana do sposobu percepcji sygnałów dźwiękowych przez człowieka. Reprezentują barwę dźwięku.

2.1.10 Chrominancja (ang. Chroma)

Wartość energii sygnału w 12 pasmach częstotliwości odpowiadających klasom wysokości dźwięku ($C, C\sharp, D, D\sharp, E, F, F\sharp, G, G\sharp, A, A\sharp, B$). Podział klas jest zgodny z systemem równomiernie temperowanym (12-TET), który jest najbardziej popularny w muzyce zachodniej.

2.2 Ramka danych

Na podstawie powyższych cech stworzono ramkę danych. Każdy wiersz reprezentuje 1 utwór o 91 atrybutach opisujących i 1 atrybucie decyzyjnym (`Genre` - gatunek muzyczny).
Atrybuty opisujące:

- `Tempo` - estymowana wartość tempa
- `STE_mean, STE_std` - średnia i odchylenie standardowe energii krótkoczasowej
- `RMS_mean, RMS_std` - średnia i odchylenie standardowe wartości skutecznej energii
- `ZCR_mean, ZCR_std` - średnia i odchylenie standardowe współczynnika przekroczeń zera
- `Centroid_mean, Centroid_std` - średnia i odchylenie standardowe centroidu widmowego
- `Bandwidth_mean, Bandwidth_std` - średnia i odchylenie standardowe szerokości pasma widma
- `Roll-off_mean, Roll-off_std` - średnia i odchylenie standardowe opadania widma
- `Constrast(0-6)_mean, Contrast(0-6)_std` - średnia i odchylenie standardowe kontrastu widmowego dla poszczególnych pasm częstotliwości
- `MFCC(0-19)_mean, MFCC(0-19)_std` - średnia i odchylenie standardowe MFCC dla poszczególnych pasm częstotliwości
- `Chroma(0-11)_mean, Chroma(0-11)_std` - średnia i odchylenie standardowe dla poszczególnych chrominancji

3 Zbiór danych

3.1 Wybrane gatunki

Celem drugiego etapu było utworzenie zbioru utworów muzycznych z 10 wybranych gatunków. Liczba utworów w każdym gatunku powinna być w miarę możliwości równa.

Wybrane gatunki muzyczne:

- muzyka chóralna a capella
- muzyka klasyczna
- muzyka elektroniczna
- muzyka ludowa
- jazz
- metal
- pop
- rap
- reggae
- rock

3.2 Ekstrakcja próbek utworów

Z każdego utworu należącego do zbioru danych wycięto zbędną ciszę z jego początku i końca oraz wyekstrahowano po 5 próbek. Próbki były kolejnymi 20-sekundowymi fragmentami utworu branymi od jego początku. Następnie zostały one poddane ekstrakcji cech, analogicznie jak dla wcześniejszego opisu. Każdej próbce przypisano etykietę `Genre` odpowiadającą jej gatunkowi muzycznemu.

3.3 Ramka danych

Wynikiem była ramka danych zawierająca 1000 obiektów o równomiernym rozkładzie klas (po 100 próbek dla każdego gatunku). Każdy obiekt ma 91 atrybutów opisujących oraz 1 atrybut decyzyjny. Utworzona ramka danych została zapisana do pliku `music_data.csv`.

4 Analiza eksploracyjna danych

W trzecim etapie przeprowadzono eksploracyjną analizę danych w celu poznania bliżej utworzonego zbioru oraz eliminacji atrybutów silnie skorelowanych.
Wszystkie atrybuty opisujące są typu zmiennoprzecinkowego, a atrybut decyzyjny typu obiektowego. Nie ma w zbiorze żadnych brakujących danych.

4.1 Statystyki rozkładu wartości atrybutów

Na podstawie analizy statystyk rozkładów wartości atrybutów wyciągnięto następujące wnioski:

- atrybuty `MFCC0_mean` i `MFCC2_mean` przyjmują wartości od rzędu -100 do rzędu 10
- atrybuty `MFCC[X]_mean` (dla wszystkich X oprócz $X \in \langle 0, 2 \rangle$) przyjmują wartości od rzędu -10 do rzędu 10
- atrybuty `STE_mean, STE_std, RMS_mean, RMS_std, ZCR_mean, ZCR_std, Chroma[X]_mean` i `Chroma[X]_std` (dla wszystkich X) przyjmują wartości rzędu 0.1
- atrybuty `Contrast0_std, Contrast1_std` i `Contrast4_std` przyjmują wartości rzędu 1
- atrybuty `Contrast2_std, Contrast3_std, Contrast5_std, Contrast6_std` i `MFCC[X]_std` (dla wszystkich X oprócz $X = 0$) przyjmują wartości od rzędu 1 do rzędu 10
- atrybuty `Contrast[X]_mean` (dla wszystkich X) przyjmują wartości rzędu 10
- atrybuty `Tempo, MFCC1_mean` i `MFCC0_std` przyjmują wartości od rzędu 10 do rzędu 100
- atrybuty `Centroid_std, Bandwidth_std` i `Roll-off_std` przyjmują wartości od rzędu 10 do rzędu 1000
- atrybuty `Centroid_mean, Bandwidth_mean` i `Roll-off_mean` przyjmują wartości od rzędu 100 do rzędu 1000

Poszczególne atrybuty mają zakresy zmienności różnych rzędów. Aby rozkłady atrybutów były do siebie zbliżone, można dokonać skalowania na dalszym etapie.

4.2 Analiza rozkładu z podziałem na klasy

Zbiór jest zbalansowany. W każdej z 10 klas występuje równa liczba obiektów. Na podstawie analizy rozkładu wartości atrybutów w poszczególnych klasach na wykresach pudełkowych wyciągnięto następujące wnioski:

- dla atrybutów `STE_mean, STE_std` oraz `RMS_mean` obiekty klas `choir` i `classical` przyjmują wartości w wąskim zakresie
- dla atrybutów `STE_mean` oraz `STE_std` obiekty klas `electronic, pop` i `rap` przyjmują wartości w szerokim zakresie
- dla atrybutów `MFCC[X]_mean` (dla $X \in \langle 4, 8 \rangle$) obiekty klasy `folk` przyjmują wartości w szerszym zakresie niż obiekty pozostałych klas

- rozkłady atrybutów `Contrast[X]_mean` (dla $X \in \langle 1, 5 \rangle$) są do siebie zbliżone
- w zbiorze występują obiekty odstające we wszystkich klasach
- nie istnieje 1 atrybut, który umożliwiłby odróżnienie od siebie wszystkich klas

4.3 Rozróżnianie klas

Przed przystąpieniem do analizy, na podstawie podobieństwa gatunków muzycznych, sformułowano hipotezy na temat klas trudnych do rozróżnienia:

- `rock` i `metal`
- `pop` i `electronic`
- `choir` i `classical`
- `rap` i `reggae`

4.3.1 Rozróżnianie klas rock i metal

Obiekty w klasach `rock` i `metal` dla większości atrybutów mają wartości w tych samych zakresach. Na przykład na podstawie atrybutów `MFCC1_std` oraz `Contrast4_mean` można odróżnić od siebie tylko część obiektów tych klas. Może to powodować pomyłki w przyszłej klasyfikacji.

4.3.2 Rozróżnianie klas pop i electronic

Dla klas `pop` i `electronic` sytuacja jest podobna jak dla `rock` i `metal`. Na przykład na podstawie atrybutów `MFCC5_std` oraz `Contrast5_mean` można odróżnić od siebie tylko małą część obiektów tych klas. Może to powodować pomyłki w przyszłej klasyfikacji.

4.3.3 Rozróżnianie klas choir i classical

Obiekty klas `choir` i `classical` przyjmują wartości w podobnych zakresach dla większości atrybutów. Jednakże całkiem dobre rozróżnienie tych klas dają atrybuty `MFCC4_std` i `MFCC5_mean`.

4.3.4 Rozróżnianie klas rap i reggae

Obiekty klasy `reggae` dla atrybutu `STE_mean` dzielą się niejako na 2. podgrupy. Jedna z nich jest dobrze rozróżnialna od klasy `rap` na podstawie atrybutów `STE_mean` i `Chroma11_mean`. Obiekty drugiej z nich przyjmują wartości w zakresie nakładającym się z zakresem wartości obiektów klasy `rap`, co uniemożliwia ich odróżnienie na tej podstawie. Obiekty klasy `rap` przyjmują wartości dla obydwu atrybutów w bardzo szerokim zakresie.

4.4 Analiza korelacji

W zbiorze danych występuje wiele skorelowanych atrybutów, zarówno pozytywnie, jak i negatywnie. Na podstawie macierzy korelacji wybrano 15 atrybutów o silnej korelacji (powyżej 0,8) do usunięcia:

- `STE_std`
- `RMS_mean`
- `Centroid_mean`
- `Centroid_std`
- `Bandwidth_mean`
- `Roll-off_mean`
- `Roll-off_std`
- `Contrast2_mean`
- `Contrast3_mean`
- `MFCC0_mean`
- `MFCC11_std`
- `MFCC13_std`
- `MFCC15_std`
- `MFCC17_std`
- `MFCC18_std`

Po usunięciu skorelowanych atrybutów, w zbiorze pozostało 76 atrybutów opisujących.

5 Klasyfikacja

5.1 Przygotowanie zbioru

W ramach czwartego etapu, zbiór danych najpierw poddano standaryzacji. Przelicznik (ang. scaler) zapisano w pliku `scaler.pkl` w celu późniejszego użycia w aplikacji.
Zbiór danych podzielono losowo na część uczącą i testową w stosunku 7:3 - 700 obiektów w zbiorze uczącym i 300 obiektów w zbiorze testowym. Losowy podział zbioru zapewnił mniej więcej równą ilość obiektów z każdej klasy w obydwu częściach.

5.2 Ocena modeli

Zastosowano 3 modele klasyfikacji:

- regresję logistyczną
- drzewo decyzyjne
- las losowy

Podczas oceny modeli brano pod uwagę 4 miary:

- dokładność (ang. accuracy) - stosunek poprawnie rozpoznanych obiektów do wszystkich obiektów
- precyzję (ang. precision) - stosunek obiektów poprawnie rozpoznanych jako klasa X do wszystkich obiektów rozpoznanych jako klasa X
- czułość (ang. recall) - stosunek obiektów poprawnie rozpoznanych jako klasa X do wszystkich obiektów klasy X
- miara F1 - średnia harmoniczna precyzji i czułości

Ponieważ badano problem klasyfikacji wieloklasowej, miary te zostały uśrednione.

5.2.1 Regresja logistyczna

W zbiorze testowym obiekty klasy `pop` były często błędnie rozpoznawane jako inne klasy (w szczególności `rap`), również wiele obiektów innych klas było błędnie rozpoznawane jako `pop` (w szczególności `electronic`). Niektóre obiekty klasy `rap` zostały błędnie rozpoznane jako `rock` oraz obiekty klasy `classical` jako `jazz`.

5.2.2 Drzewo decyzyjne

Przeanalizowawszy wartości dokładności, precyzji, czułości i miary F1 dla zbioru testowego, wybrałam drzewo decyzyjne o głębokości = 17. Dalsze pogłębianie drzewa nie ma wpływu na wartości badanych miar.
W zbiorze testowym obiekty klasy `metal` były często błędnie rozpoznawane jako `rock`, obiekty klasy `electronic` jako `pop`, obiekty klasy `classical` jako `jazz`, obiekty klasy `rap` jako `electronic` oraz obiekty klasy `reggae` jako `rap`. Wiele z tych pomyłek pokrywa się z postawionymi wcześniej hipotezami na temat podobieństwa i rozróżnialności klas.

5.2.3 Las losowy

Przeanalizowawszy wartości dokładności, precyzji, czułości i miary F1 dla zbioru testowego, wybrałam las losowy o liczbie drzew = 400. Dalsze zwiększanie liczby drzew pogarsza wartości badanych miar.

W zbiorze testowym wiele obiektów (ok. 1/4) klasy `rock` zostało błędnie rozpoznanych jako `metal`. Niektóre obiekty klasy `pop` zostały błędnie rozpoznane jako `electronic` oraz obiekty klasy `reggae` jako `rock`.

5.3 Porównanie modeli

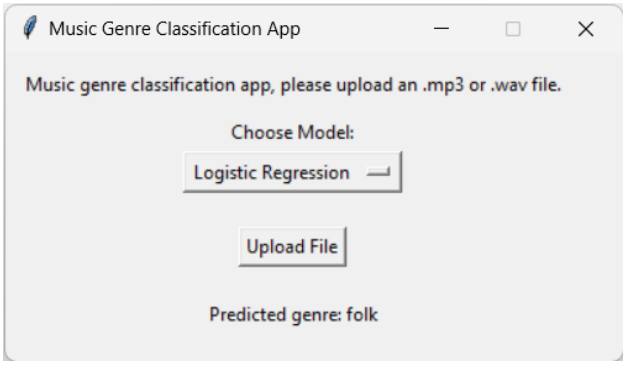
Ostatecznie osiągnięto następujące wartości miar dotyczących klasyfikacji obiektów zbioru testowego dla poszczególnych modeli:

	dokładność	precyzja	czułość	miara f1
regresja logistyczna	0,74	0,76	0,74	0,74
drzewo decyzyjne	0,60	0,62	0,60	0,60
las losowy	0,82	0,83	0,82	0,82

Średnio, dla zbioru testowego, las losowy cechował się najwyższymi wartościami miar dokładności, precyzji, czułości i miary f1. Najgorzej wypadło drzewo decyzyjne.

6 Aplikacja okienkowa

Zaimplementowano prostą aplikację okienkową umożliwiającą klasyfikację utworu wgranego przez użytkownika jako plik dźwiękowy. Możliwe jest wgranie pliku w formacie `.mp3` lub `.wav`. Przed wgraniem pliku należy wybrać 1 z 3 dostępnych modeli z listy rozwijanej (`Logistic Regression, Decision Tree` lub `Random Forest`). Z wgranego pliku wycinane jest 5 20-sekundowych próbek (po kolei, od początku utworu) i ekstrahowane są cechy (76 ostatecznie wybranych do klasyfikacji na etapie analizy eksploracyjnej danych). W wyniku tego powstaje ramka danych o rozmiarze 5 x 76. Za pomocą przelicznika dopasowanego do zbioru uczącego i testowego, cechy są standaryzowane. Następnie ramka przekazywana jest do wybranego modelu. Każda z 5 próbek otrzymuje etykietę (gatunek). Jako ostateczna etykieta wybierany jest gatunek o największej liczbie wystąpień (głosowanie większościowe). W przypadku takiej samej liczby głosów na więcej niż 1 gatunek, etykieta wybierana jest według porządku alfabetycznego. Czas od wybrania pliku do wyświetlenia przewidzianego gatunku wynosi ok. 6 sekund.



7 Wnioski

7.1 Rozłączność gatunków

Klasyfikacja gatunków muzycznych jest kłopotliwa, ponieważ wiele z nich ma cechy wspólne. Nie da się precyzyjnie zdefiniować, gdzie leży granica między 2 gatunkami, np. `rock` i `metal`. Dlatego większość utworów na rynku muzycznym jest hybrydą wielu gatunków, zawiera w sobie elementy charakterystyczne dla kilku na raz. Często w życiu codziennym po przesłuchaniu utworu mam problem w zidentyfikowaniu, jaki gatunek sobą reprezentuje, co przekładało się na trudność późniejszego stwierdzenia poprawności klasyfikacji.

7.2 Utworzenie zbiór danych

Utworzenie zbioru danych własnoręcznie nie było łatwym zadaniem. Miałam trudności w wyborze utworów, które "dobrze" reprezentują dany gatunek. Ponieważ nawet w obrębie 1 gatunku muzycznego jest duża różnorodność, zbiór danych powinien był być większy. Uważam, że ekstrakcja kilku próbek z każdego utworu nie była dobrym pomysłem. Sprawiało to, że w zbiorze uczącym i testowym mogły występować niemalże identyczne obiekty (próbki sąsiednie z tego samego utworu), co zaprzecza idei podziału na zbiór uczący i testowy. Wyniki miar jakości klasyfikacji dla zbioru testowego nie reprezentowały prawdziwie zdolności generalizacji modelu. Ponieważ dobór parametrów drzewa decyzyjnego oraz lasu losowego bazowałam właśnie na wielkościach tych miar, modele były skłonne do przetrenowania. Uzyskiwały bardzo dobre wyniki dla utworów będących częścią oryginalnego zbioru danych, natomiast z innymi (nowymi) utworami już sobie nie radziły tak dobrze.

8 Repozytorium

Projekt został udokumentowany w repozytorium dostępnym pod adresem: `https://github.com/lamachan/music_genre_classification`.

Pliki w repozytorium:

- dane - katalog z plikami tekstowymi zawierającymi zbiór danych:
  - `music_data_v1.csv` - plik tekstowy ze zbiorem danych po etapie ekstrakcji danych
  - `music_data_v2.csv` - plik tekstowy ze zbiorem danych po etapie eksploracyjnej analizy danych
  - `music_data_v3.csv` - plik tekstowy ze zbiorem danych po standaryzacji

- `librosa_testing.ipynb` - testowanie funkcji biblioteki `librosa`
- `feature_extraction.ipynb` - ekstrakcja danych
- `data_mining.ipynb` - eksploracyjna analiza danych
- `classification_models.ipynb` - uczenie i ocena modeli
- `classification_app` - katalog z plikami dotyczącymi aplikacji okienkowej
  - `pickles` - katalog z plikami utrwalającymi typu `.pkl`:
    - `scaler.pkl` - przelicznik dokonujący standaryzacji
    - `decision_tree_model.pkl` - model drzewa decyzyjnego
    - `logistic_regression_model.pkl` - model regresji logistycznej
    - `random_forest_model.pkl` - model lasu losowego
  - `feature_extraction.py` - moduł zawierający funkcje ekstrahujące cechy z pliku dźwiękowego
  - `app.py` - logika aplikacji i GUI
  - `classification_app_gui.png` - zrzut ekranu GUI aplikacji

9 Źródła

- dokumentacja pakietu `librosa`
- P. Rao, Audio Signal Processing Chapter in Speech, Audio, Image and Biomedical Signal Processing using Neural Networks, (Eds.) Bhanu Prasad and S. R. Mahadeva Prasanna, Springer-Verlag, 2007
- musicinformationretrieval.com
- https://towardsdatascience.com/are-you-dropping-too-many-correlated-features-d1c96654abe6
- https://towardsdatascience.com/comprehensive-guide-on-multiclass-classification-metrics-af94cfb83fbd