Attachment 13

Rozszerzone podsumowanie w języku polskim rozprawy pod tytułem „Algorytmiczne metody wykrywania tempa i sygnatur czasowych utworów muzycznych"

Jeremiah O. Abimbola
2024

Wstęp

Metrum, często przedstawiane jako ułamek na początku partytury, oznacza strukturę rytmiczną utworu muzycznego. Wskazuje on liczbę uderzeń w każdym takcie (licznik) i rodzaj nuty tworzącej jedno uderzenie (mianownik). Na przykład metrum 4/4 wskazuje cztery uderzenia na takt, a ćwierćnuta otrzymuje jedno uderzenie.
W niniejszej rozprawie wskazano kilka wyzwań związanych z tym zadaniem, m.in.:

Różnice rytmiczne i synkopy: Obecność różnorodnych wzorców rytmicznych, nieoczekiwanych akcentów i odchyleń od tradycyjnych struktur metrycznych, szczególnie w gatunkach takich jak jazz czy muzyka latynoska, komplikuje proces identyfikacji spójnej sygnatury czasowej.

Wiele sygnatur czasowych w kompozycji: Dokładne wykrywanie i śledzenie zmian sygnatur czasowych, które mogą być nagłe lub stopniowe, stanowi poważne wyzwanie, zwłaszcza w złożonych aranżacjach muzycznych.

Polirytmy: Współistnienie wielu wzorców rytmicznych jednocześnie, powszechnych w muzyce afrykańskiej lub afro-kubańskiej, wymaga algorytmów zdolnych do dokonania rozplotu i analizy tych nakładających się struktur.

Ornamenty muzyczne i ekspresyjne wyczucie czasu: Wykonawcy często wprowadzają zmiany tempa i czasu za pomocą technik takich jak „accelerando" (stopniowy wzrost tempa), co utrudnia rozpoznanie podstawowej sygnatury czasu.

Cel pracy dyplomowej

W niniejszej pracy wykazano wyższą wydajność modeli uczenia głębokiego w klasyfikowaniu metrum z sygnałów audio. W przypadku klasyfikacji binarnej przy użyciu cech MFCC, modele uczenia głębokiego osiągnęły dokładność od 88% (CRNN) do 89% (CNN i ResNet18-LSTM), podczas gdy w przypadku klasyfikacji binarnej przy użyciu cech spektrogramu, dokładność wahała się od 98% (CNN) do 99,67% (ResNet18). W przypadku klasyfikacji wieloklasowej przy użyciu cech MFCC, modele uczenia głębokiego osiągnęły dokładność od 80,29% (CRNN) do 86% (ResNet18-LSTM), a dla klasyfikacji wieloklasowej przy użyciu cech spektrogramu, dokładność wahała się od 82,86% (CNN) do 90,71% (ResNet18).

Zakres rozprawy

Analizowane zadanie ma istotne zastosowanie w wydobywaniu informacji z muzyki (ang. Music Information Retrieval, MIR), zwłaszcza, że współczesna konsumpcja muzyki w coraz większym stopniu opiera się na sygnałach audio, a nie na sygnałach MIDI.

Klasyczne cyfrowe metody przetwarzania sygnału (ang. Digital Signal Processing, DSP)

Badanie rozpoczyna się kompleksowym przeglądem klasycznych metod DSP do wykrywania sygnatur czasowych, które obejmują:

- **Matryca podobieństwa audio (ASM):** Ta metoda identyfikuje powtarzające się takty w utworze przez porównanie dłuższych segmentów audio (taktów) z krótszymi fragmentami. Konstruuje macierze podobieństwa na podstawie spektrogramu w celu uchwycenia powtarzających się wzorców. Ocena zbioru danych Meter2800 dała dokładność na poziomie 53% w przypadku klasyfikacji binarnej i 51% w przypadku klasyfikacji wieloklasowej.

- **Matryca podobieństwa uderzeń (BSM):** Technika ta opiera się na śledzeniu uderzeń i analizie spektrogramu, dzieląc spektrogram na klatki synchroniczne z rytmem. BSM generuje macierz podobieństwa w oparciu o korelację krzyżową między uderzeniami, określając licznik na podstawie analizy przekątnej tej macierzy. BSM osiągnął dokładność 50% w przypadku klasyfikacji binarnej i 49% w przypadku klasyfikacji wieloklasowej w zestawie danych Meter2800.

- **Matryca podobieństwa współczynników cepstralnych częstotliwości Mel (MFCCSM):** Proponowana jako nowatorskie podejście, MFCCSM wykorzystuje współczynniki cepstralne częstotliwości Mel (MFCC) do przechwytywania informacji barwowych i przedstawiania tekstur muzycznych. Generuje macierz podobieństwa na podstawie spektrogramu Mel, analizując jego przekątną w celu określenia metrum utworu. MFCCSM uzyskało lepsze wyniki niż ASM i BSM z dokładnością 55% w przypadku klasyfikacji binarnej i 53% w przypadku klasyfikacji wieloklasowej.

Pomimo swojej użyteczności, klasyczne metody DSP często mają trudności z uchwyceniem niuansów i złożoności występujących w rzeczywistych sygnałach audio, co powoduje potrzebę stosowania bardziej wyrafinowanych podejść.

Zbiór danych Meter2800

Aby wytrenować i ocenić proponowane modele, wprowadzono nowy zbiór danych o nazwie Meter2800. Ten zbiór danych składa się z 2800 próbek audio z adnotacjami, starannie wybranych z ustalonych zbiorów danych muzycznych, takich jak GTZAN, FMA-medium i MagnaTagATune, podzielonych na cztery klasy z różnym metrum. Meter2800 niweluje krytyczną lukę w istniejących zbiorach danych, w których często brakuje metrum z adnotacjami ekspertów dla sygnałów audio, opierając się zamiast tego na potencjalnie niedokładnych szacunkach. Utworzenie tego zbioru danych ma kluczowe znaczenie dla postępu badań nad wykrywaniem metrum przez dostarczanie wysokiej jakości danych do celów treningu i oceny.

Algorytmy uczenia maszynowego (ML)

Przechodząc do algorytmów uczenia maszynowego, pod kątem wykrywania metrum ocenianych jest kilka klasycznych metod ML:

- **Maszyna wektorów nośnych (SVM):** Osiągnęła najwyższą dokładność wśród modeli ML, osiągając 86,67% w przypadku klasyfikacji binarnej i 74,29% w przypadku klasyfikacji wieloklasowej.

- **K-najbliższych sąsiadów (KNN):** Wykazano konkurencyjną w stosunku do SVM skuteczność z dokładnością 84,83% w przypadku klasyfikacji binarnej i 70,43% w przypadku klasyfikacji wieloklasowej.

- **Naiwny klasyfikator Bayesa:** Uzyskał wyniki porównywalne z SVM - 84,50% dla klasyfikacji binarnej i 66,00% dla klasyfikacji wieloklasowej.

- **Las losowy:** Również wykazał się dobrą skutecznością, osiągając dokładność 86,00% w przypadku klasyfikacji binarnej i 72,57% w przypadku klasyfikacji wieloklasowej.

Te modele uczenia maszynowego przetestowano na zestawie danych Meter2800, który zapewnia solidną podstawę do oceny ich wydajności w wykrywaniu metrum.

Modele głębokiego uczenia się (DL).

W badaniu szczegółowo zbadano modele uczenia głębokiego, które wykazały doskonałą wydajność w przypadku złożonych zadań rozpoznawania wzorców:

- Konwolucyjna sieć neuronowa (CNN)
- Konwolucyjno-rekurencyjna sieć neuronowa (CRNN)
- ResNet18
- ResNet18-LSTM

Te modele uczenia głębokiego konsekwentnie przewyższały zarówno klasyczne modele DSP, jak i ML, osiągając dokładność w zakresie od 88,00% do 99,67% w przypadku zadań klasyfikacji binarnej.

Techniki optymalizacji

Aby zwiększyć wydajność modelu, zastosowane zostały techniki optymalizacji, takie jak algorytmy genetyczne (GA) i optymalizacja Bayesa. GA wykorzystano do optymalizacji wag przypisanych do różnych współczynników MFCC, natomiast optymalizacja Bayesa pomogła określić optymalne parametry zarówno dla modelu MFCCSM, jak i GA. Zoptymalizowany model MFCCSM wykazał znaczną poprawę, osiągając średnią dokładność na poziomie 63,5%.

Zespołowe metody uczenia się

Wreszcie, w badaniu zbadano techniki uczenia się zespołowego łącząc predykcje z wielu modeli przy użyciu różnych strategii głosowania, w tym większościowego głosowania i Bayesowskiego modelu uśredniającego (ang. Bayesian Model Averaging, BMA). Metody zespołowe wykazały dalszą poprawę dokładności zarówno w wykrywaniu sygnatur czasowych binarnych, jak i wieloklasowych podczas integracji klasycznych modeli DSP, ML i DL. GA został również użyty do optymalizacji wag w jednej z technik głosowania w modelach zepsołowych (weighted voting).

Podsumowując, niniejsza rozprawa znacząco rozwija dziedzinę wykrywania metrum w sygnałach audio, przeprowadzając kompleksowy przegląd istniejących metod klasyfikacji muzyki, proponując nowy model (MFCCSM) w celu poprawy dokładności wykrywania metrum, tworząc zbiór danych Meter2800 specjalnie zaprojektowany do oceny metrum przez algorytmy wykrywania sygnatur na danych audio, badając różne algorytmy uczenia maszynowego i głębokiego, zoptymalizowane przy użyciu optymalizacji Bayesa i GA oraz wykazanie, że połączenie tradycyjnych modeli z zaawansowanymi technikami może znacznie

zwiększyć skuteczność wykrywania metrum. Badania te kładą mocny fundament pod przyszłe prace w ramach MIR, oferując obiecujące możliwości dalszego rozwoju i zastosowań w przetwarzaniu sygnałów audio, jednocześnie wnosząc pozytywny wkład do społeczności przez publicznie dostępne repozytoria kodów.