

Zadanie nr 2 - Wyznaczanie cech oraz badanie trendu

Analiza danych złożonych z detekcją wyjątków

Karol Kazusek - 254189, Sebastian Zych - 254264

21.10.2024

1 Cel zadania

Zadanie polega na wyznaczeniu cech w wybranym strumieniu danych oraz wyznaczeniu trendu. Do analiz należało wybrać tematykę medycyny lub rozpoznawania faz ruchu aktywności człowieka (systemy HAR)

2 Opis zaproponowanych Klasyfikatorów i algorytmów detekcji zmian

2.1 Metoda klasyfikacyjne

2.1.1 K-nn

Metoda **K-Najbliższych Sąsiadów (K-NN)** to jedna z najprostszych technik klasyfikacji. Działa na zasadzie porównywania nowego punktu danych z istniejącymi przykładami treningowymi. Algorytm wybiera K najbliższych sąsiadów (na podstawie odległości euklidesowej lub innych miar), a następnie przypisuje nowy punkt do klasy, która występuje najczęściej wśród tych sąsiadów.

- **n_neighbors**: Liczba sąsiadów (K) do uwzględnienia przy klasyfikacji.
- **weights**: Sposób ważenia sąsiadów. Możliwe wartości to 'uniform' (wszyscy sąsiedzi mają taką samą wagę) i 'distance' (bliźsi sąsiedzi mają większy wpływ).
- **metric**: Miara odległości używana do porównania punktów, w naszym przypadku wyznaczamy odległość za pomocą metryki wyznaczoną przez DTW.

2.1.2 DTW

Metryka **Dynamic Time Warping (DTW)** to algorytm służący do mierzenia odległości między dwoma sekwencjami czasowymi, które mogą różnić się częstotliwością, fazą lub długością. Jego celem jest znalezienie optymalnej nieliniowej ścieżki dopasowania między dwiema sekwencjami, minimalizując różnice między nimi. Algorytm wykorzystuje dynamiczne programowanie, aby efektywnie znaleźć najlepsze dopasowanie.

2.2 Metoda detekcji zmian

2.2.1 ADWIN

(Okno przesuwne) **Adaptive Windowing (ADWIN)** to metoda wykrywania concept drift, czyli zjawiska zmiany rozkładu danych w czasie, co może wpływać na skuteczność modeli uczących się. ADWIN działa poprzez dynamiczne dostosowywanie okna czasowego, które monitoruje dane strumieniowe, starając się zidentyfikować momenty, w których statystyczne właściwości danych ulegają

zmianie. Kiedy wykrywana jest istotna różnica między starą a nową częścią okna, starsze dane są odrzucane, a nowe włączane do analizy. Dzięki automatycznej adaptacji wielkości okna ADWIN skutecznie reaguje na różne rodzaje i prędkości driftu, co sprawia, że jest użytecznym narzędziem w środowiskach strumieniowych.

- **delta:**

- Typ: `float`
- Opis: Kontroluje poziom istotności statystycznej przy wykrywaniu driftu. Niższa wartość oznacza bardziej restrykcyjne wykrywanie zmian, wyższa – częstsze wykrywanie.

- **clock:**

- Typ: `int`
- Opis: Określa, co ile próbek algorytm sprawdza możliwość wystąpienia driftu.

- **min_window_length:**

- Typ: `int`
- Opis: Minimalna długość okna obserwacyjnego, chroni przed zbyt szybkim usuwaniem starszych danych.

- **grace_period:**

- Typ: `int`
- Opis: Minimalna liczba obserwacji przed rozpoczęciem testu na concept drift.

3 Charakterystyka wybranych do danych

MIT-BIH Arrhythmia Database to szeroko stosowany zestaw danych wykorzystywany w badaniach dotyczących analizy sygnałów elektrokardiograficznych (EKG) oraz automatycznej klasyfikacji arytmii. Zbiór ten zawiera 48 zapisów sygnałów EKG zarejestrowanych u 47 pacjentów, w tym zarówno osób z różnymi typami arytmii, jak i zdrowych. Każdy zapis obejmuje około 30 minut sygnału EKG, który został pobrany w częstotliwości 125 Hz.

Dane zostały ręcznie oznakowane przez kardiologów, co pozwala na identyfikację różnych typów arytmii, takich jak np. skurcze dodatkowe, migotanie przedsionków czy blokady serca. MIT-BIH Arrhythmia Database jest często wykorzystywany w algorytmach sztucznej inteligencji do trenowania systemów do automatycznej diagnozy arytmii.

Kroki używane do ekstrakcji uderzeń z sygnału EKG według autorów DBLP:journals/corr/abs-1805-0079DBLP:journals/corr/abs-1805-007944 były następujące:

1. Podział ciągłego sygnału EKG na okna 10s i wybór jednego okna 10s z sygnału EKG.

2. Normalizacja wartości amplitudy do zakresu od zera do jeden.
3. Znalezienie zbioru wszystkich lokalnych maksimów na podstawie zerokrosów pierwszej pochodnej.
4. Znalezienie zbioru kandydatów na szczyty R EKG poprzez zastosowanie progu 0.9 na znormalizowanej wartości lokalnych maksimów.
5. Znalezienie mediany interwałów czasowych R-R jako nominalnego okresu bicia serca dla danego okna (T).
6. Dla każdego szczytu R wybór części sygnału o długości równej 1.2T.
7. Uzupełnienie każdej wybranej części zerami, aby jej długość była równa zdefiniowanej stałej długości.

Liczba próbek: 21900

Liczba kategorii: 5

Częstotliwość próbkowania: 125Hz

Nazwa klasy	etykieta	liczebność	
Normalne uderzenie	0	1658	
Przedwczesne uderzenie nadkomorowe	1	556	Dane w celu trenowania
Przedwczesny skurcz komorowy	2	1448	
Fuzja skurczu komorowego i normalnego uderzenia	3	162	
Niezdyscyplinowane uderzenie	4	1608	

nia klasyfikatorów 1-NN-DTW i 3-NN-DTW zostały okrojone do liczebności 100 na każdą klasę, przez wysokie wymagania sprzętowe algorytmów.

4 Eksperymenty i wyniki

4.1 Eksperyment nr 1

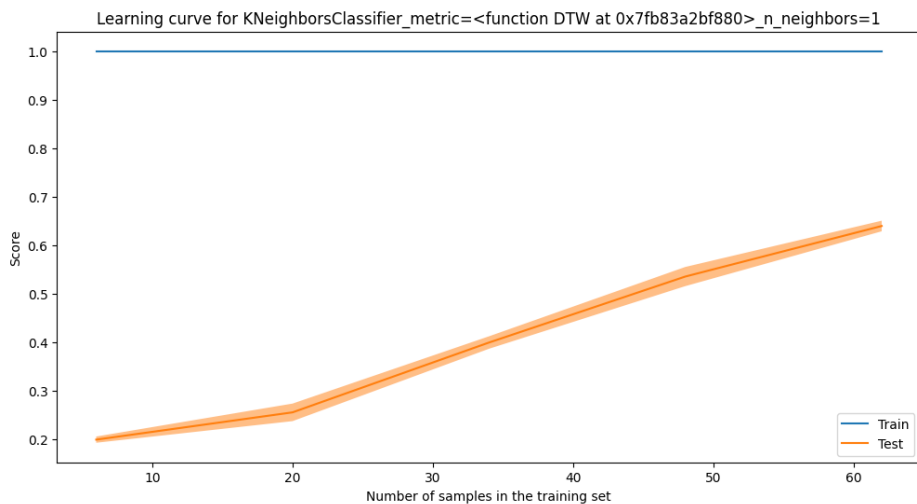
4.1.1 Założenia

Wykonanie klasyfikatora z metryką DTW dla różnych liczby sąsiadów.

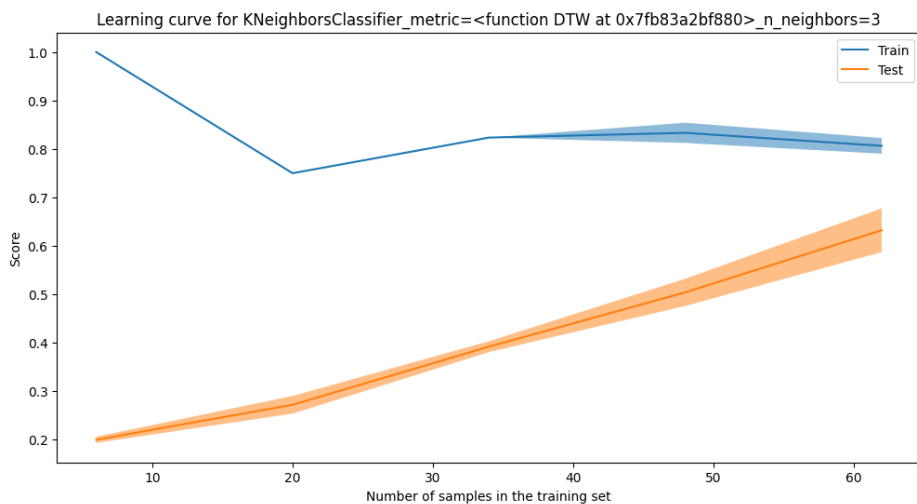
4.1.2 Rezultat 1-NN-DTW i 3-NN-DTW

Classifier	Accuracy	Precision	Sensitivity	Specificity
KNeighborsClassifier_metric=DTW_n_neighbors=1	0.660	0.682765	0.660	0.915
KNeighborsClassifier_metric=DTW_n_neighbors=3	0.620	0.622787	0.620	0.905

Tabela 1: Statystyki dla 1-NN-DTW i 3-NN-DTW

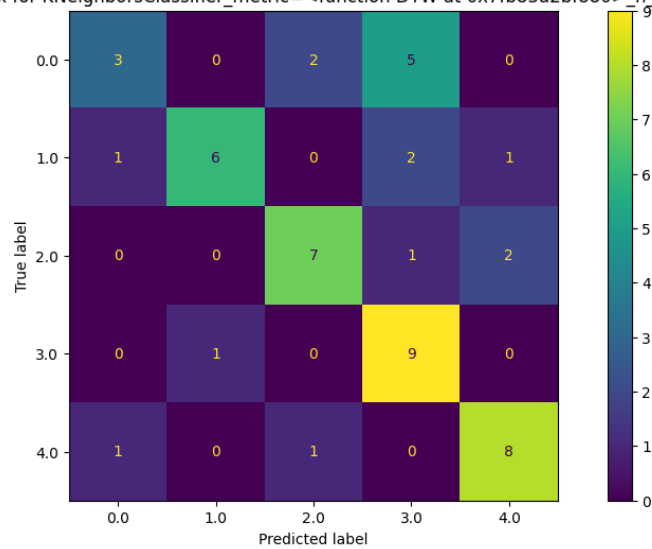


Rysunek 1: Krzywa uczenia dla 1-NN-DTW



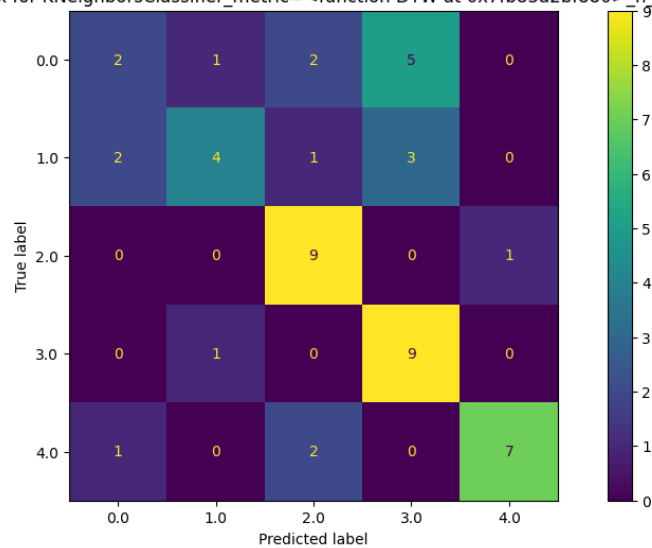
Rysunek 2: Krzywa uczenia dla 3-NN-DTW

Confusion Matrix for KNeighborsClassifier_metric=<function DTW at 0x7fb83a2bf880>_n_neighbors=1



Rysunek 3: Macierz pomylek dla 1-NN-DTW

Confusion Matrix for KNeighborsClassifier_metric=<function DTW at 0x7fb83a2bf880>_n_neighbors=3



Rysunek 4: Macierz pomylek dla 3-NN-DTW

4.2 Eksperyment nr 2

Wykorzystania algorytmu detekcji zmian ADWIN w celu poznania jego działania oraz lepszego poznania zbioru danych.

4.2.1 Rezultat ADWIN

Numer indeksu wykrycia zmiany	Na jaką klasę
Change detected at index 1695	1
Change detected at index 2239	2
Change detected at index 3711	3
Change detected at index 3871	4

Tabela 2: Miejsce i klasa zmiany trendu

5 Wnioski

Wnioski z przeprowadzonych eksperymentów dowodzą, że

- Efektywność klasyfikatorów wykorzystujących z miary DTW jest niższa w przypadkach większych zbiorów danych posiadających więcej cech.
- Utrudnione jest trenowanie klasyfikatorów wykorzystujących miarę DTW, ponieważ liczenie tej miary jest wymagającym zadaniem dla procesora przez swoją złożoną implementację wykorzystuje programowanie dynamiczne.
- Metody detekcji dryfu są w stanie z pewną pewnością określić **concept drift**, w przypadku **ADWIN** wykorzystywane jest okno przesuwne.
- Metody detekcji concept drift umożliwiają lepiej poznać zbiór danych strumieniowych oraz mogą być wykorzystywane jako triggerzy poszczególnych operacji w przypadku zmiany trendu.