Warsztaty z technik uczenia maszynowego

Wojciech Klusek, Aleksander Kuś

24 May 2022

1 Wstęp

Celem projektu było zaimplementowanie algorytmu ROCKET (RandOm Convolutional KErnel Transform) w języku R dla szeregów wielu zmiennych przy założeniu, że obserwacje mogą mieć różne długości. Algorytm ROCKET służy do klasyfikacji szeregów czasowych z wykorzystaniem dużych ilości losowych "kerneli", które mają parametry takie jak: "length", "weights, "bias", "dilation" oraz "padding".

2 Przygotowanie danych

Wstępne dane zostały pobrane ze strony timeseriesclassification.com. Do testów użyliśmy zestawów danych BasicMotions, Epilepsy, Handwriting oraz NATOP. Następnie pobrane dane przycięliśmy tak, aby obserwacje miały różne długości zgodnie z następującymi zasadami:

Dla każdej klasy

- 1/3 instancji będzie miała długość [10%,40%] oryginalnej długości
- 1/3 instancji będzie miała długość (40%,70%] oryginalnej długości
- 1/3 instancji będzie miała długość (70%,100%] oryginalnej długości

Skrócone instancje zostały następnie wypełnione średnią arytmetyczną wartości, które nie zostały przycięte dla danej instancji lub zerami.

3 Opis algorytmu ROCKET

Na wstępie generowana jest zadana ilość "kerneli" z odpowiednimi parametrami. Następnie są one aplikowane do danych treningowych oraz testowych. "Kernel" ma następujące parametry:

- "Length" - losowo wybrana liczba ze zbioru { 7, 9, 11 }
- "Weights" - wartości z rozkładu normalnego $X \sim \mathcal{N}(0, 1)$
- "Bias" - wartość z rozkładu jednostajnego ciągłego $\mathcal{U}(0, 1)$
- "Dilation" - jest próbkowana w skali wykładniczej $d = \lfloor 2^x \rfloor$, $x \sim \mathcal{U}(0, A)$, gdzie $A = \log_2 \frac{l_input-1}{l_kernel-1}$
- "Padding" - w momencie generacji "kernela" podejmowana jest losowa decyzja czy "padding" ma zostać użyty w momencie aplikowania "kernela" czy też nie. Bez paddingu "kernele" nie są wyśrodkowane w pierwszych i ostatnich $\lfloor l_{kernel} - 1 \rfloor$ punktów.

Każdy "kernel" (ω) z "Dilation" (d) jest aplikowany do każdego wejściowego szeregu czasowego (X) od pozycji i, według następującego wzoru:

$$X_i * \omega = \sum_{j=0}^{l_{kernel}-1} X_{i+(j \times d)} \times \omega_j.$$

Rocket oblicza dwie zagregowane cechy z każdej mapy cech, tworząc dwie liczby rzeczywiste dla każdego "kernela":

- maksymalna wartość
- odsetek wartości dodatnich (ppv)

4 Uruchomienie programu

Do uruchomienia programu wymagane są następujące biblioteki:

Dla R

- foreign
- reticulate

Dla pakietu reticulate wymagana jest instalacja programu "miniconda". Pakiet ten wykorzystywany jest dla funkcji "array_reshape()", której odpowiednika nie znaleźliśmy w czystym języku R.

Instalacja powyższych zależności:

```
install.packages(c("foreign", "reticulate"))
library("reticulate")
install_miniconda()
```

Dla Pythona

- sktime
- numpy

Z biblioteki sktime wykorzystywana jest klasa RidgeClassifierCV używana do oceniania poprawności algorytmu. Do uruchomienia projektu użyliśmy środowiska PyCharm Community, uruchamiając plik run.py.

5 Przeprowadzone testy

Dla każdego wyżej opisanego zbioru danych przycięliśmy zbiory "train" i "test" według schematu opisanego powyżej oraz uruchomiliśmy nasz algorytm. Następnie wyniki przekazaliśmy do klasyfikatora w celu oceny. Proces ten powtórzyliśmy 10 razy z uwagi na brak determinizmu etapu przycinania danych, a otrzymane wyniki uśredniliśmy. Wyniki przedstawiono poniżej.

6 Wyniki

Wyniki zostały wyznaczone dla 100 "kerneli"

BasicMotions

Parametry zbioru danych:

- długość instancji TRAIN: 40
- długość instancji TEST: 40
- Długość szeregów: 100

Numer próby	Wynik dla średniej	Wynik dla zer	Wynik referencyjny
1	0.875	0.875	0.9
2	0.9	0.95	0.9
3	0.9	0.9	0.9
4	0.9	0.875	0.9
5	0.95	0.95	0.9
6	0.875	0.9	0.9
7	0.9	0.9	0.9
8	0.875	0.9	0.9
9	0.975	0.975	0.9
10	0.975	0.975	0.9
Średnia	0.9125	0.920	0.9

Epilepsy

Parametry zbioru danych:

- długość instancji TRAIN: 137
- długość instancji TEST: 138
- Długość szeregów: 206

Numer próby	Wynik dla średniej	Wynik dla zer	Wynik referencyjny
1	0.913	0.891	0.942
2	0.876	0.876	0.942
3	0.884	0.876	0.942
4	0.905	0.891	0.942
5	0.898	0.905	0.942
6	0.876	0.847	0.942
7	0.905	0.891	0.942
8	0.891	0.884	0.942
9	0.855	0.833	0.942
10	0.855	0.876	0.942
Średnia	0.886	0.877	0.942

Handwriting

Parametry zbioru danych:

- długość instancji TRAIN: 150
- długość instancji TEST: 850
- Długość szeregów: 152

Numer próby	Wynik dla średniej	Wynik dla zer	Wynik referencyjny
1	0.191	0.197	0.248
2	0.192	0.191	0.248
3	0.197	0.196	0.248
4	0.192	0.197	0.248
5	0.198	0.202	0.248
6	0.182	0.192	0.248
7	0.201	0.202	0.248
8	0.202	0.204	0.248
9	0.187	0.187	0.248
10	0.190	0.195	0.248
Średnia	0.193	0.196	0.248

NATOPS

Parametry zbioru danych:

- długość instancji TRAIN: 180
- długość instancji TEST: 180
- Długość szeregów: 51

Numer próby	Wynik dla średniej	Wynik dla zer	Wynik referencyjny
1	0.605	0.605	0.827
2	0.611	0.661	0.827
3	0.605	0.627	0.827
4	0.577	0.638	0.827
5	0.594	0.65	0.827
6	0.594	0.661	0.827
7	0.594	0.65	0.827
8	0.633	0.633	0.827
9	0.6	0.577	0.827
10	0.6	0.661	0.827
Średnia	0.601	0.636	0.827

7 Wnioski

Z przeprowadzonych eksperymentów wynika, że dokładność algorytmu dla danych nieobciętych jest największa. Obcinanie danych i wypełnianie obciętych wartości zerami daje lepsze rezultaty niż wypełnianie wartością średnią pozostałych elementów. Dla zbioru "Handwriting", dla którego ilość instancji w zbiorze treningowym jest dużo mniejsza od tych w zbiorze testowym, dokładność była najmniejsza. Obcinanie danych ma różny wpływ na dokładność w różnych zbiorach. W zbiorze NATOPS, gdzie długość szeregów była najmniejsza, obcinanie danych miało największy wpływ.

8 Bibliografia

[1] https://github.com/alan-turing-institute/sktime/blob/main/sktime/transformations/panel/rocket/_rocket.py

[2] https://github.com/alan-turing-institute/sktime/blob/main/examples/rocket.ipynb

[3] https://github.com/angus924/rocket