

Wydział Elektroniki i Technik Informacyjnych
Politechnika Warszawska

Zaawansowane zagadnienia sieci neuronowych

Predykcja liczby zachorowań na COVID-19 za pomocą grafowych sieci
czasowo-przestrzennych - dokumentacja wstępna

Belniak Michał,
Młyniec Paweł

Warszawa, 14 kwietnia 2021

1 Opis zadania

Zgodnie z informacjami na stronie projektu, zadanie polega na zaimplementowaniu i wytrenowaniu modelu czasowo-przestrzennej grafowej splotowej sieci neuronowej do predykcji liczby zachorowań na COVID-19 dla danych ze Stanów Zjednoczonych. Do tego celu należy posłużyć się danymi dostarczonymi przez The New York Times [2], które są podzielone na dane dla całego państwa, dla stanów i dla hrabstw. W naszym przypadku interesują nas dane z podziałem na hrabstwa. Zadanie ma obejmować zweryfikowanie działania modelu na danych benchmarkowych o zachorowaniach na ospę wietrzną w Węgrzech w latach 2004-2014 [3] oraz porównanie modelu z prostymi lokalnymi modelami autoregresyjnymi.

Wykorzystywane przez nas dane są od siebie zależne zarówno w wymiarze czasowym jak i wymiarze przestrzennym, stąd wykorzystanie grafowej sieci czasowo-przestrzennych ma swoje uzasadnienie. Każde hrabstwo może być reprezentowane przez wierzchołek grafu przechowujący liczbę zachorowań w tym hrabstwie w danym momencie, a krawędzie grafu mogą łączyć wierzchołki odpowiadające hrabstwom sąsiadującym ze sobą lub blisko położonym, lecz niekoniecznie ze sobą sąsiadującym. Pewnym czynionym w ten sposób założeniem, tudzież testowaną hipotezą, jest fakt zależności liczby zachorowań w danym hrabstwie od liczby zachorowań w hrabstwach sąsiadujących.

2 Sposób rozwiązania problemu i wyjaśnienie poczynionych założeń

2.1 Typ modelu

Naszym modelem będzie grafowa splotowa sieć czasowo-przestrzenna [1]. Jest to model sieci stworzony aby mierzyć się z problemami czasowo przestrzennymi, w których występują zależności między danymi zarówno w wymiarze czasowym jak i przestrzennym. Oryginalny model na którym będziemy bazować został wykorzystany do predykcji ruchu drogowego w sieci ciągów komunikacyjnych. Zamiast zwyczajnych jednostek konwolucyjnych oraz rekurencyjnych, aplikowany jest problem na grafie i stosowana jest pełna struktura konwolucyjna. Model ten pasuje dobrze do problemu zadania, jako że rozprzestrzenianie się COVID-19 jest zależne od czasu i hipotetycznie od przestrzennego rozmieszczenia ludności oraz poziomu przepływu ludności między lokalizacjami.

Stworzenie grafu będzie wymagało uzyskania współrzędnych hrabstw by wyznaczyć odległość między nimi. Jako że problematyczne może być automatyczne określenie sąsiedztwa hrabstw, wykorzystamy pomysł z oryginalnego artykułu opisującego model [1], gdzie tworzenie macierzy sąsiedztwa odbywało się na podstawie odległości między hrabstwami według wzoru:

$$w_{ij} = \begin{cases} \exp\left(-\frac{d_{ij}^2}{\sigma^2}\right) & \text{if } i \neq j \text{ i } \exp\left(-\frac{d_{ij}^2}{\sigma^2}\right) > \epsilon \\ 0 & \text{w p. p.} \end{cases}, \quad (1)$$

gdzie w_{ij} jest wyznaczoną wagą krawędzi (wagą sąsiedztwa) uwarunkowaną przez odległość d_{ij} między hrabstwem i oraz j . σ^2 oraz ϵ są stałymi kontrolującymi rozkład i gęstość macierzy sąsiedztwa z oryginalnymi wartościami odpowiednio 10 i 0.5. Przy takich wartościach tych stałych wartość w_{ij} jest różna od 0 dla odległości mniejszych niż 2.63, więc w przypadku przyjęcia jednostki d jako kilometry z pewnością trzeba będzie dostosować wartości σ^2 oraz ϵ , aby macierz nie była zbyt rzadka. Najprawdopodobniej będzie to edytowalnym parametrem algorytmu.

2.2 Wybrane zbiory danych

Tak jak wspomniano w opisie zadania wykorzystywany będzie zbiór danych dotyczący dziennych zachorowań na COVID-19 w Stanach Zjednoczonych [2] oraz dane benchmarkowe o zachorowaniach na ospę wietrzną w Węgrzech [3].

2.3 Narzędzia

Do rozwiązania problemu posłużymy się biblioteką PyTorch Geometric Temporal [4]. Zawiera ona różne dynamiczne oraz temporalne geometryczne głębokie sieci neuronowe w tym grafowe sieci czasowo-przestrzenne.

Kod będzie się uruchamiał po postawieniu środowiska *conda* i będzie ustrukturyzowany z wykorzystaniem biblioteki *PyTorchLightning*.

Do wizualizacji prawdopodobnie użyjemy biblioteki Streamlit.

2.4 Zakładana funkcjonalność

Nasz projekt będzie realizował predykcję zachorowań na COVID-19 oraz na ospę wietrzną. Po nauczaniu modelu dane będą wizualizowane na mapie. Przewidujemy możliwość przeglądania różnic pomiędzy danymi rzeczywistymi oraz predykcją.

3 Bibliografia

- [1] *Spatio-temporal Graph Convolutional Neural Network: A Deep Learning Framework for Traffic Forecasting*, Bing Yu and Haoteng Yin and Zhanxing Zhu, 2017
- [2] *Coronavirus (Covid-19) Data in the United States*, The New York Times, (Dostęp zdalny 14.04.2021) <https://github.com/nytimes/covid-19-data>
- [3] *PyTorch Geometric Temporal Dataset* PyTorch, (Dostęp zdalny 14.04.2021) <https://pytorch-geometric-temporal.readthedocs.io/en/latest/modules/dataset.html>
- [4] *pytorch_geometric_temporal* PyTorch, (Dostęp zdalny 14.04.2021) https://github.com/benedekrozemberczki/pytorch_geometric_temporal
- [5] *Streamlit* (Dostęp zdalny 14.04.2021) <https://docs.streamlit.io/en/stable/api.html#streamlit.map>