Wykonali:
Paweł Gędłek
Patryk Wójtowicz

Wizualizacja dużych zbiorów danych

Raport z postępów w projekcie:
k-NN Sampling for Visualization of Dynamic data using LION-tSNE

Artykuł dotyczący zagadnień poruszanych w projekcie:
https://ieeexplore.ieee.org/abstract/document/8990391

Repozytorium z kodem LION tSNE:
https://github.com/andreyboytsov/LION-tSNE

Jupyter Notebook z przeprowadzonymi testami (w załączniku do raportu)

1. Wstęp.

Na początku projektu zdecydowaliśmy się na analizę działania metody LION tSNE (Local Interpolation with Outlier coNtrol t-distributed stochastic neighbor embedding) i zwracanych przez niego rezultatów. W pierwszej kolejności zajęliśmy się sprawdzeniem jaki wpływ na wizualizację ma odpowiednie próbkowanie danych wejściowych. Serię eksperymentów przeprowadziliśmy na zbiorach IRIS oraz MNIST, a jako próbkę kontrolną wybraliśmy tradycyjne tSNE. W połowie eksperymentów mieliśmy do czynienia z losowo wybranym zbiorem testowym natomiast w drugiej połowie posłużyliśmy kNN samplingiem będącym jedną z części naszego projektu.

2. Sampling.

Losowe próbkowanie polega na pseudolosowym doborze rekordów z wybranego datasetu, co jest obarczone możliwością wyboru nierównych podzbiorów danych klas oraz ryzykiem dużego stopnia wariancji danych.

```
mnist_random_df['target'].value_counts()
```

```
7    230
1    228
9    216
3    206
5    193
0    193
8    190
6    188
2    186
4    170
Name: target, dtype: int64
```

Podział na poszczególne klasy dla zbioru MNIST i losowego sposobu próbkowania

Obecnie kNN sampling przeprowadzamy być może dość naiwną metodą jednak zwracającą póki co obiecujące wyniki, a mianowicie wyliczamy za pomocą Nearest Centroid Classifier pochodzącego z biblioteki scikit-learn centroidy dla poszczególnych klas. Następnie dla każdej klasy wyszukujemy k najbliższych sąsiadów centroida klasy i umieszczamy je w zbiorze testowym. Tak przeprowadzone próbkowanie rozwiązuje oba problemy wspomniane powyżej (występujące w losowym próbkowaniu).

```
mnist_knn_df['target'].value_counts()
```

```
2    200
0    200
3    200
6    200
9    200
1    200
8    200
7    200
4    200
5    200
Name: target, dtype: int64
```

Podział na poszczególne klasy dla zbioru MNIST i próbkowania k najbliższych sąsiadów

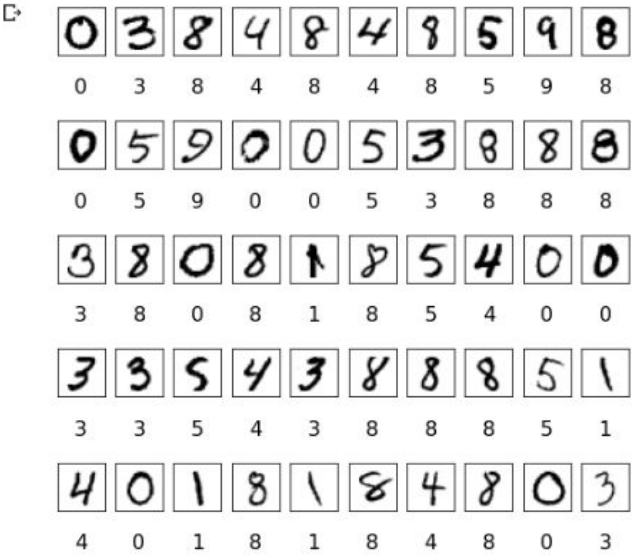
Kolejnym krokiem, który chcemy wykonać jest próbkowanie w dynamicznie zmieniającym się zbiorze danych co stanowi główny powód, dla którego warto stosować metodę LION tSNE zamiast metody tSNE.

3. Wyniki przeprowadzonych eksperymentów.

W prezentowanych eksperymentach po lewej znajduje się wynik wizualizacji z losowo wybranymi próbkami, natomiast po prawej statyczny kNN sampling.

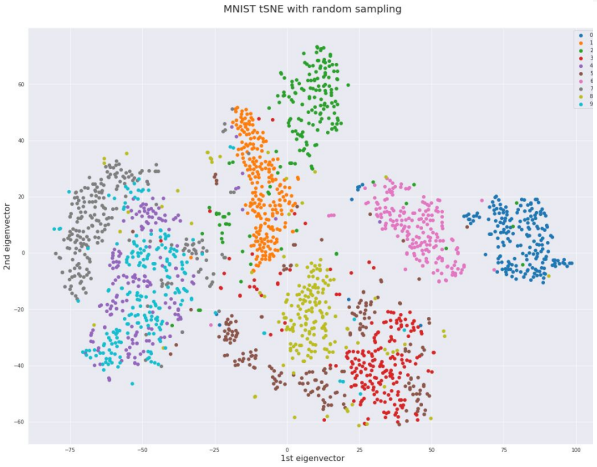
a) MNIST DATASET

Dataset zawierający opisy cyfr pisanych reprezentowanych jako obrazki 28x28 pikseli. W przypadku eksperymentów na tym zbiorze danych posłużyliśmy zbiorem testowym zawierającym 2000 rekordów (mniej więcej 200 obrazków na daną klasę).

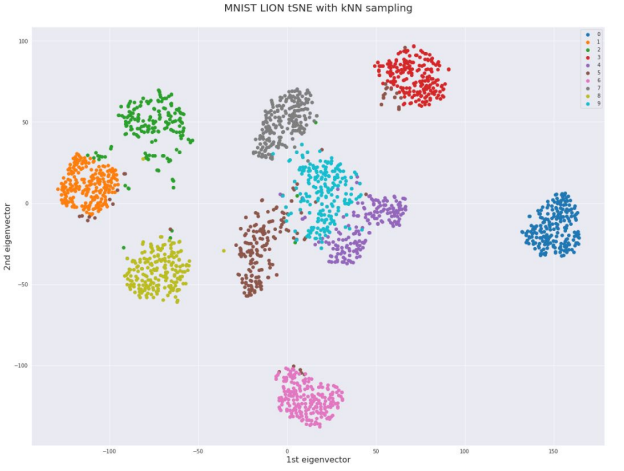


Wizualizacja przykładowych danych ze zbioru MNIST

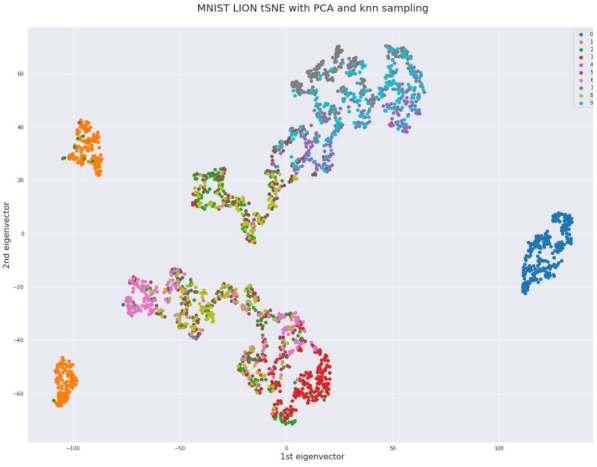
- t-SNE

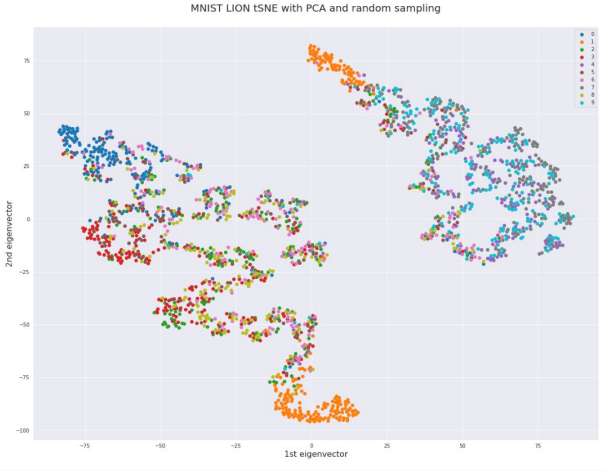


- LION t-SNE



- LION tSNE z PCA

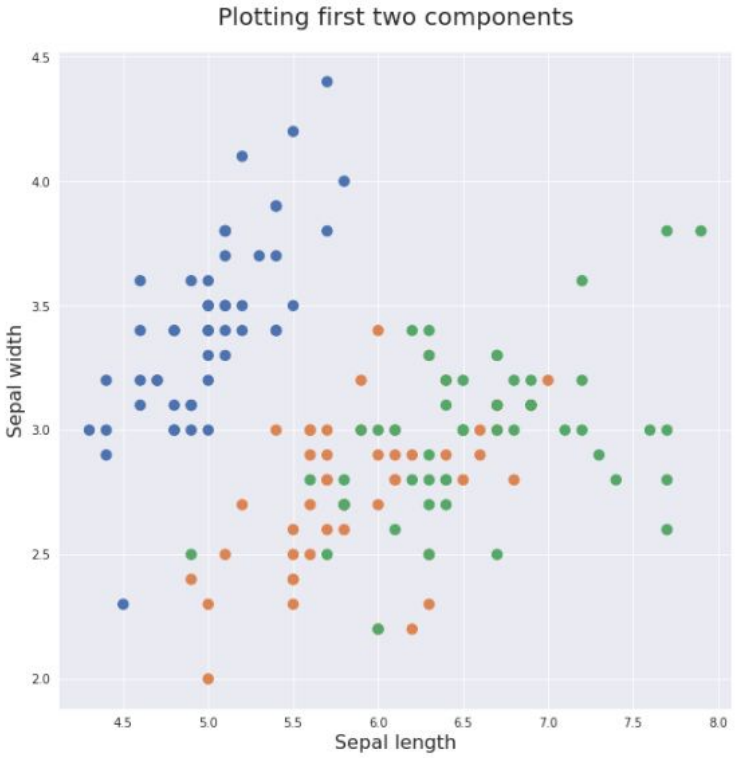




- LION tSNE z MDS

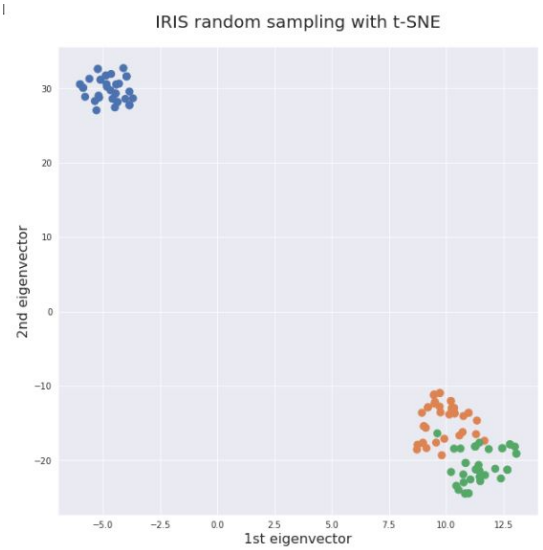
b) IRIS DATASET

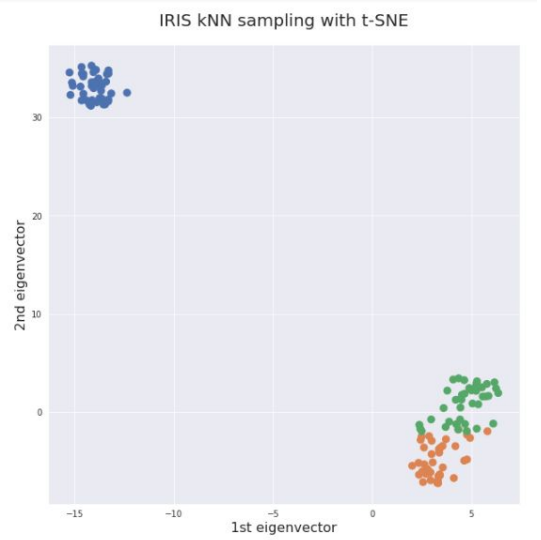
Jest to stosunkowo mały zbiór zawierający opisy czterech własności, trzech różnych gatunków irysów. Z racji na niewielki rozmiar, zbiór okazał się przydatny do testowania samplingu danych (wybieraliśmy 120 spośród 150 dostępnych rekordów), jednak wydaję się zbyt mały dla niektórych metod, aby stworzyć w pełni wiarygodny model.



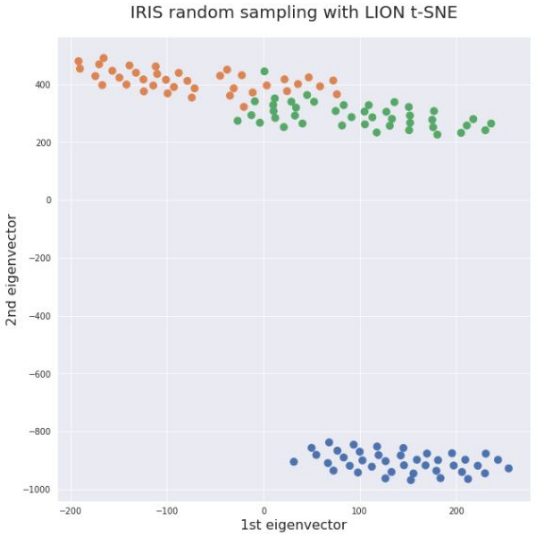
Wizualizacja przykładowych danych ze zbioru IRIS

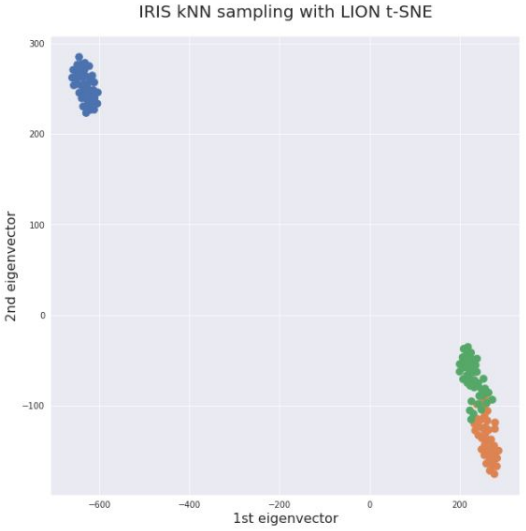
- t-SNE



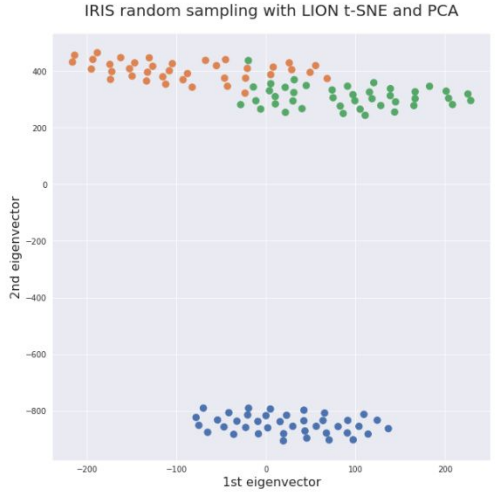


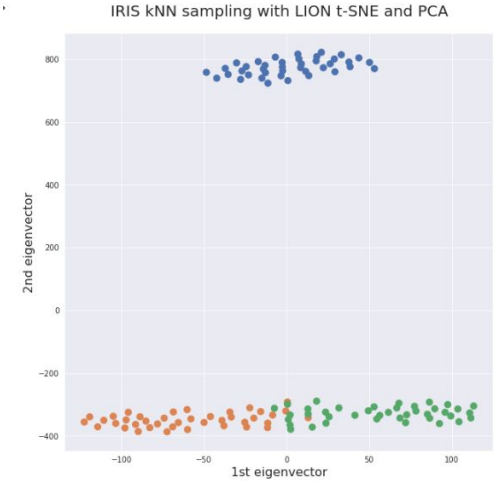
- LION t-SNE



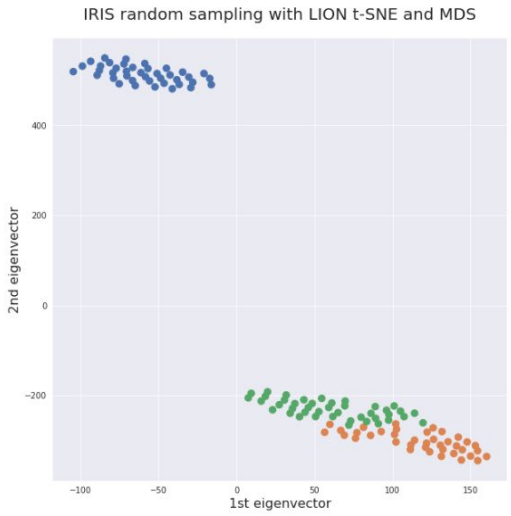


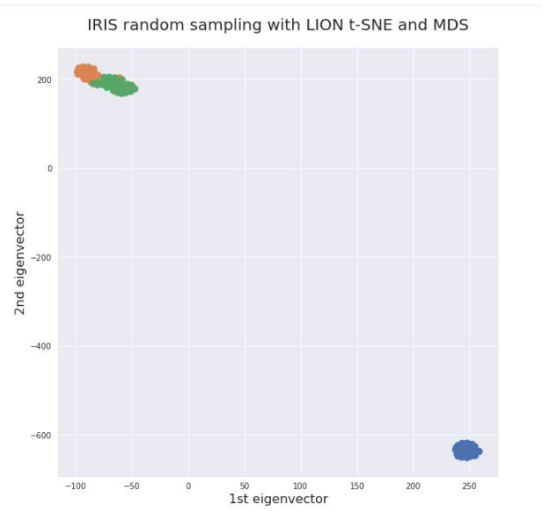
- LION t-SNE z PCA





- LION t-SNE z MDS





4. Wnioski i dalsze plany pracy nad projektem.

Podstawowym wnioskiem płynącym z tej części projektu jest zasadność użycia odpowiedniego próbkowania danych wejściowych, aby uzyskać lepsze wyniki wizualizacji, co widać na załączonych diagramach pochodzących z eksperymentów.

W kolejnym etapie zamierzamy zaimplementować dynamiczne próbkowanie (wykorzystywane na przykład w medycynie, gdy aktualne dane zmieniają się na bieżąco) i sprawdzić, jak radzi sobie z nim tradycyjne tSNE oraz jego modyfikacja LION tSNE. Chcielibyśmy także podjąć się wizualizacji rozmieszczenie Outlierów oraz innych elementów ważnych z punktu widzenia działania metody tSNE oraz jej poszczególnych modyfikacji przedstawionych w artykule.