Nauczanie maszynowe – projekt zaliczeniowy

Autor:

Michał Szczygieł

Projekt został podzielony na dwie części o podobnej tematyce ale o różnej problematyce. Taki podział z założenia powinien dostarczyć okazji do sprawdzenia w praktyce różnych metod/algorytmów nauczania maszynowego. W projekcie wykorzystane zostały algorytmy takie jak: PCA, KMeans i wielowarstwowe modele neuronowe. Eksplorowane są także różne podejścia do reprezentacji danych pod konkretne algorytmy.

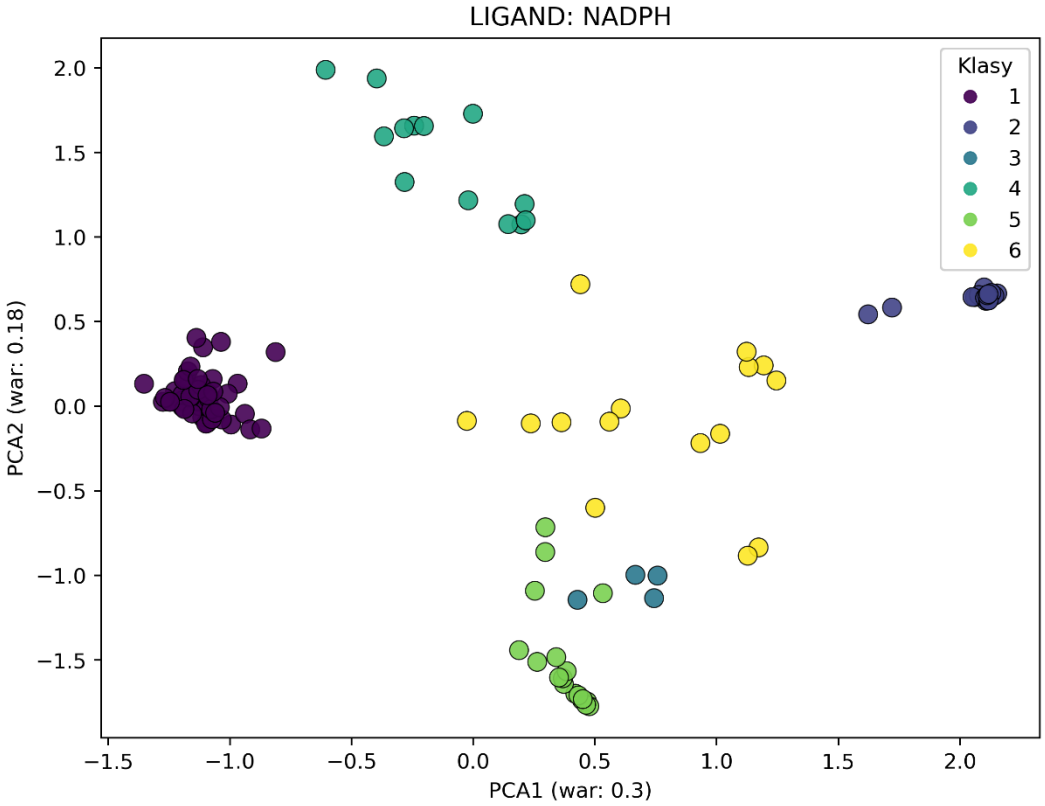
Tematyka projektu:

W ramach projektu badane są możliwości wykorzystania algorytmów uczenia maszynowego do eksploracji danych o kieszeniach (białkowych) wiążących ligandy. Dane te są pozyskiwane z plików PDB (dostępnych w bazie danych biologicznych RCSB PDB). Struktury krystaliczne (białek i związanych przez nie ligandów) są reprezentowane w postaci diagramów Woronoja. Taka reprezentacja pozwala na łatwe uzyskanie informacji o sąsiedztwie wybranych grup atomów (reszt, ligandów). Wyznaczenie miejsc wiązania ligandów w białkach odbywa się poprzez znalezienie komórek diagramu Woronoja rozpiętych nad atomami liganda. Reprezentacja tak znalezionych miejsc wiązania jest zależna od wybranego algorytmu (wektoryzacja, one-hot).

Grupowanie miejsc wiążących ligandy – badanie zróżnicowania miejsc wiążących na przykładzie struktur krystalicznych:

Rozwiązanie tego problemu składa się z następujących kroków (w kolejności): wyznaczenie miejsc wiążących jeden konkretny ligand np. cholesterol, wektoryzacja miejsc wiązania, redukcja wymiarowości na powstałym zbiorze wektorów, grupowanie w przestrzeni zredukowanej. O tym jak wyznaczane są miejsca wiązania wspomniane zostało w poprzedniej sekcji. Tworzenie reprezentacji wektorowej polega na przypisaniu każdemu atomowi liganda wektora embeddingowego zawierającego informację o otoczeniu. Następnie

wszystkie te wektory są łączone w jeden, reprezentujący całe miejsce wiązania. Kolejnym krokiem jest redukcja wymiarowości. Ta przeprowadzana jest na zbiorze zwektoryzowanych miejsc wiążących z wykorzystaniem algorytmu PCA przy zachowaniu ~80% wariancji. Na końcu wykonywane jest grupowanie algorytmem KMeans w przestrzeni zredukowanej. Poniżej znajduje się przykład grupowania miejsc wiążących NADPH (przeprowadzono również dla CLR i ATP):



Widoczna jest wyraźna separacja miejsc wiązania już w dwóch pierwszych wymiarach przestrzeni zredukowanej.

Więcej przykładów znajduje się zeszycie Jupyter (*grouping.ipynb*). Szczegóły techniczne zostały wyjaśnione w komentarzach kodu.

Przewidywanie liganda pasującego do kieszeni wiążącej (one-hot):

W drugiej części projektu badana jest możliwość przewidywania ligandów pasujących do miejsc wiązania. W pierwszym podejściu opis miejsca wiązania odbywa się za pomocą 20D wektora 0-1 reprezentującego obecność lub brak aminokwasu w składzie kieszeni wiążącej. W celu rozwiązania postawionego

problemu skorzystano z sieci neuronowych. Ich zadaniem było przewidzenie liganda pasującego do konkretnych wektorów one-hot miejsc wiązania.

O danych:

- 6 klas ligandów (szczegóły w zeszycie),

- po 100 miejsc wiązania na ligand,

- podział na zbiory uczący i treningowy 80:20

Wyniki:

Przetestowano kilka architektur sieci. W przypadku każdej osiągnięto dokładność w zbiorze testowym na poziomie ~80% (przewidywanie 6 klas).

Szczegóły (architektury sieci, wykresy uczenia) znajdują się w zeszycie Jupyter (*one_hot_ml.ipynb*).

Przewidywanie liganda pasującego do kieszeni wiążącej (embeddingi):

To jest drugie podejście do poprzedniego problemu. Jedyną różnicą jest zmiana sposobu reprezentacji miejsc wiązania. Zamiast reprezentacji one-hot wykorzystywana jest reprezentacja wektorowa – tak jak w problemie grupowania. Na wejście sieci podawane są wektory w przestrzeni zredukowanej (20D).

Wyniki:

Podobnie jak w poprzednim kodowaniu osiągane dokładności oscylują w okolicy 80% (przewidywanie 6 klas).

Szczegóły dostępne zeszycie Jupyter (*vector_ml.ipynb*).

Podsumowanie:

Udało się z powodzeniem zrealizować postawione cele. Uzyskane wyniki są satysfakcjonujące. Ten projekt nie wyczerpuje przedstawionej problematyki. Przedstawione sposoby reprezentacji miejsc wiązania można dalej modyfikować

lub tworzyć nowe – skuteczniejsze (np. wokselizacja kieszeni wiążących i wykorzystanie sieci z konwolucją w 3D).