Multi Omics Data - Raport

Jakub Giezgała

21 czerwca 2025

Z bazy *OncoKB*^{TM} pobrano 125 genów z listy *Vogelstein et al. (2013)* oraz zidentyfikowano 48 912 rekordów związanych z mutacjami somatycznymi. Odpowiednie pliki zapisano w formacie `.parquet` z wykorzystaniem biblioteki *PyArrow*. Następnie sprawdzono, czy zbiór etykiet `primary_site` jest niezbalansowany, poprzez przedstawienie histogramu. Przykładowo, nowotwór pęcherzyka żółciowego reprezentowany jest przez jedynie dwie próbki, podczas gdy nowotwór macicy – przez ponad 100 tysięcy.

Następnie dokonano ekstrakcji motywu 3-nukleotydowego, znormalizowanego względem pirymidyn (C lub T), co daje łącznie 96 możliwych wariantów. Dotyczy to mutacji typu SNP, a także DEL oraz INS, dla których również uwzględniono sąsiedztwo nukleotydów. Dla nich jednak sprawdzano, czy długość mutacji wpływa na zmianę ramki odczytu (czy jest podzielna przez 3), co dodało kolejne 64 motywy. Alternatywnym podejściem byłaby tokenizacja każdego motywu mutacyjnego, jednak prowadziłoby to do zbyt dużej liczby wariantów, skutkując nadmiernym dopasowaniem modelu.

Każdej mutacji przypisano bin (przedział) co 100 000 nukleotydów, na podstawie pozycji początkowej i końcowej oraz przyporządkowano odpowiedni chromosom. Następnie połączono zmienne: `bin`, `gene`, `motif` (wszystkie zakodowane binarnie) oraz zmienną ciągłą `rna`, którą poddano normalizacji log1p, bazując na wartościach FPKM z górnego kwantyla.

Ze względu na silne niezbalansowanie klas w zmiennej `primary_site`, zdecydowano się na odrzucenie przypadków będących singletonami. Zbiór danych podzielono proporcjonalnie na 70/10/20 – odpowiednio dla zbiorów treningowego, walidacyjnego i testowego. Uwzględniono również odpowiednie wagi dla każdej klasy.

Zaprojektowano podstawową architekturę modelu, w której dla każdej zmiennej przewidziano jedną warstwę ukrytą typu `dense`, a następnie cztery warstwy łączące informacje z różnych źródeł. Do wyboru najlepszego modelu wykorzystano wartość `accuracy` dla zbioru walidacyjnego, przy użyciu funkcji straty `CrossEntropyLoss`, uwzględniającej niezbalansowanie klas. Oprócz tego monitorowano miary: F1-score, Recall oraz Precision – liczone bez odpowiednich wag.

W kolejnym kroku wykorzystano bardziej zaawansowaną architekturę – z mniejszą liczbą warstw, regularyzacją grzbietową oraz mechanizmem `dropout`. Liczbę neuronów dla każdej grupy zmiennych dostosowano metodą bisekcji, a także zmniejszono wartość współczynnika uczenia (*learning rate*).

Dla modelu bazowego, pomimo przyzwoitego F1-score na poziomie 0,73, wartość accuracy z uwzględnieniem niezbalansowanych klas wynosiła jedynie 0,51. Z przebiegu procesu uczenia widać również, że wartość funkcji straty (loss) dla zbioru walidacyjnego rosła, mimo jej spadku dla zbioru treningowego – co jest klasycznym przykładem nadmiernego dopasowania (overfittingu). Ulepszona architektura nie wykazywała tego problemu – strata się stabilizowała, a model osiągał wyniki o ponad 10 punktów procentowych lepsze. Co istotne, model był w stanie nauczyć się predykcji również dla mniej licznych klas.