POLITECHNIKA ŚLĄSKA

KATEDRA INŻYNIERII I BIOLOGII SYSTEMÓW

Rozprawa doktorska

Opracowanie nowych algorytmów uczenia maszynowego dla heterogenicznych danych biomedycznych

Autor: mgr inż. Agata Wilk

Promotor: prof. dr hab. inż. Krzysztof Fujarewicz

Gliwice, czerwiec 2024

Streszczenie

Hodowle komórek zdesynchronizowanych w fazie cyklu komórkowego, tkanki nowotworowe złożone z komórek różnych typów i o różnych profilach molekularnych, kohorty pacjentów różniących się stanem zaawansowania choroby, czynnikami genetycznymi i środowiskowymi, zróżnicowane populacje regionów czy krajów. Dane kliniczne, genomiczne, transkryptomiczne, proteomiczne i obrazowe, pochodzące z różnych źródeł, o różnych typach, strukturze i wymiarowości. Heterogeniczność jest nieodłącznym aspektem badań biomedycznych, przyczyną wielu zjawisk biologicznych i podstawą personalizowanej terapii. Analiza heterogenicznych danych prezentuje jednak szereg wyzwań — od złego uwarunkowania numerycznego i utrudnionej estymacji parametrów modelu, do obecności różnej liczby wektorów cech dla poszczególnych obiektów.

Kluczowym problemem naukowym, którego rozwiązanie jest przedmiotem niniejszej rozprawy, jest wykorzystanie danych dla heterogenicznych obiektów w uczeniu maszynowym. Pracę doktorską stanowi cykl siedmiu artykułów naukowych, obrazujących własne doświadczenia autorki.

Pierwsza część poświęcona jest heterogeniczności występującej na różnych poziomach. W szczególności omówiony jest dylemat pomiędzy budowaniem wspólnego modelu dla heterogenicznych grup, a niezależną analizą podgrup czy pojedynczych obiektów. Jako kompromis łączący te strategie, zaproponowano autorskie, indywidualizowane podejście do estymacji parametrów (na przykładzie modelu epidemiologicznego), polegające na estymacji części parametrów jako wspólnych dla wszystkich obiektów, a części jako niezależnych. Zastosowanie indywidualizowanego modelowania pomogło przezwyciężyć problem złego uwarunkowania numerycznego, pozwalając jednocześnie zachować elementy indywidualnej charakterystyki, co znalazło odzwierciedlenie w rozkładzie błędów dopasowania.

Druga część rozprawy porusza problem heterogeniczności struktury danych, czyli obecności różnej liczby wektorów cech. Przedstawiono oryginalne strategie wykorzystania takich danych w klasyfikacji (dla obrazowania proteomicznego) oraz modelowaniu przeżycia (dla wielu gromadzeń w obrazowaniu PET/CT) oparte na agregacji wektorów cech lub agregacji wyników modelowania. Otrzymane wyniki pokazują, że wykorzystanie informacji pochodzącej ze wszystkich dostępnych wektorów cech poprzez zastosowanie agregacji pozwala na poprawę zdolności predykcyjnej modeli względem wykorzystania pojedynczego wektora cech.

Słowa kluczowe: heterogeniczne dane, uczenie maszynowe, klasyfikacja, indywidualizowany model, analiza przeżycia, agregacja