STRESZCZENIE

Rozprawa koncentruje się na rozwoju narzędzi informatycznych służących do identyfikacji biomarkerów chorobowych oraz wspomagania diagnostyki chorób poprzez analizę danych multiomicznych z wykorzystaniem metod statystycznych i technik uczenia maszynowego. Motywacją do podjęcia badań w tym obszarze były wyzwania zidentyfikowane w dwóch projektach będących częścią badań klinicznych, EPISTOP i EPIMARKER, których celem było opracowanie modeli predykcyjnych umożliwiających przewidywanie wystąpienia napadów padaczkowych lub lekooporności u dzieci ze stwardnieniem guzowatym oraz identyfikację biomarkerów procesu chorobowego. W wyniku tej analizy wyłoniono zagadnienia wymagające uzupełnienia i generalizacji, co stanowiło podstawę do rozszerzenia zakresu badań z wykorzystaniem danych z otwartych baz danych multiomicznych. Pogłębione badania dotyczyły problemu selekcji cech, gdzie przeanalizowano wpływ różnych metod selekcji na wyniki klasyfikacji oraz opracowano zestaw procedur wspierających użytkownika w wyborze odpowiedniej metody. Następnie przeprowadzono badania nad zastosowaniem algorytmów uczenia zespołowego do rozwiązania problemu braków w danych multiomicznych, analizując dwie metody selekcji modeli pod kątem ich zdolności adaptacji do niekompletnych danych. Zaproponowano modyfikację miary niezgodności dla par klasyfikatorów uwzględniającą obserwacje niesklasyfikowane przez wcześniejsze modele z grupy. Tak zdefiniowana metryka wykazała zdolność do selekcji niewielkiej liczby modeli, które umożliwiają poprawną klasyfikację większości obserwacji, w których występują częściowe braki danych. Wyniki realizacji projektów EPISTOP, EPIMARKER oraz badań nad problemem selekcji cech i uczenia zespołowgo zostały wykorzystane w konstrukcji potoku przetwarzania playOmics. W rozprawie opisano szczegóły implementacji, przykładowe zastosowanie oraz porównanie wyników działania potoku do innych metod. Potok playOmics umożliwia uproszczenie zarządzania różnorodnymi danymi omicznymi, przetwarzanie wstępne, budowę i ocenę modeli klasyfikacyjnych oraz identyfikację biomarkerów, jednocześnie dostarczając narzędzia zwiększające interpretowalność analizy i ułatwiające jej odtwarzalność. W ostatniej części rozprawy zaprezentowano implementację opracowanego potoku playOmics w badaniu klinicznym DIPGen, umożliwiającą binarną klasyfikację pacjentów dla zdefiniowanego przez użytkownika celu analizy, na podstawie danych pochodzących z różnych warstw omicznych.

Słowa kluczowe: uczenie maszynowe, metody statystyczne, dane multiomiczne, poszukiwanie biomarkerów, diagnostyka chorób, choroby rzadkie