Politechnika Śląska
Katedra Inżynierii i Biologii Systemów

Alicja Płuciennik

Algorytmy integracji danych i uczenia maszynowego w diagnostyce nowotworów

Promotor: prof. dr hab. inż. Krzysztof Fujarewicz
Promotor pomocniczy: dr hab. Małgorzata Oczko–Wojciechowska
Opiekun przemysłowy: dr inż. Michał Bachorz

Gliwice 2021

Streszczenie pracy doktorskiej

mgr inż. Alicja Płuciennik

Algorytmy integracji danych i uczenia maszynowego w diagnostyce nowotworów

Diagnostyka nowotworów jest jednym z istotnych zagadnień współczesnej medycyny, ponieważ dotyczą one coraz większej liczby osób. Szybka diagnostyka zmian złośliwych i określenie podtypu nowotworu wpływają na dalsze decyzje dotyczące terapii. Popularnym kierunkiem rozwoju diagnostyki onkologicznej są badania molekularne, jednak ich dostępność w rutynowej diagnostyce nie jest wystarczająca, co często tłumaczone jest wysokim kosztem badania wykorzystującego wielogenowe klasyfikatory molekularne.

Celem tej pracy jest przedstawienie przykładów i strategii zastosowania algorytmów integracji danych do tworzenia narzędzi diagnostycznych, wykorzystujących dane molekularne oraz inne czynniki kliniczne zastosowane w celu redukcji liczby cech molekularnych i zmniejszenia kosztu badania.

W pracy szczegółowo przeanalizowano zastosowanie fuzji danych oznaczające połączenie zbiorów cech dla każdej obserwacji podlegającej klasyfikacji. Praca obejmuje zbadanie efektów zastosowania dwóch strategii fuzji danych, wyróżnionych w kontekście selekcji cech jako wczesna fuzja danych oraz późna fuzja danych, na jakość klasyfikatora - wyrażoną jako dokładność klasyfikacji, czyli odsetek poprawnie sklasyfikowanych próbek. Ponieważ praca ma charakter wdrożeniowy, badania wykonano na zbiorach rzeczywistych danych pochodzących z zebranych próbek guzów tarczycy. Dla luminalnych podtypów raka piersi wykorzystano dane pobrane z bazy TCGA. Ze względu na warunki przeprowadzenia badań wynikające ze stosunkowo nielicznych zbiorów, analizy wykonano z użyciem metody bootstrap. Dodatkowo, efekty fuzji danych zbadano w kontekście użycia metod ekstrakcji cech klinicznych lub bez ekstrakcji. Szczegółowo zbadano zależności między cechami poddanymi fuzji danych korzystając ze wskaźników takich jak korelacja Spearman'a oraz informacja wzajemna. Jako propozycję porównania wpływu fuzji na selekcję cech zaproponowano analizę stabilności selekcji cech z wykorzystaniem indeksu Kunchevy.

Wyniki przeprowadzonych badan wskazują, że zastosowanie fuzji danych umożliwia redukcję liczby cech molekularnych, jednak wyniki powiązane są z występowaniem zależności pomiędzy cechami. Wykazano statystycznie istotne różnice (p-wartość < 0.05) na korzyć zastosowania fuzji danych w przypadku guzów tarczycy. Podsumowując przeprowadzone analizy, można stwierdzić, że do późnej fuzji danych można wykorzystać cechy wykazujące same w sobie dobrą jakość klasyfikacji i posiadające zależności z cechami w zbiorze podstawowym. Jeżeli jednak nie jesteśmy pewni co do wystarczającej siły zależności oraz wystarczającej jakości klasyfikacji dla dodawanych cech, korzystne jest zastosowanie wczesnej fuzji, tak aby algorytm selekcji cech wybrał najlepsze zmienne z podzbioru. Z przedstawionych wyników analizy zależności między cechami wynika, że aby fuzja była skuteczna nie muszą to być silne zależności - wystarczy średnia korelacja oraz informacja wzajemna mniejsza od 0.4.

Z racji wdrożeniowego charakteru pracy przedstawiono również przykładowe wyniki z badań przemysłowych ilustrujące zastosowanie przedstawionych metod fuzji danych dla zbiorów danych nowotworów tarczycy oraz raków piersi. Przedstawione wyniki pracy oraz metody i algorytmy postępowania stanowią przedstawienie możliwości stosowania fuzji danych oraz mogą być zastosowane w przyszłości w szerszym kontekście wykorzystując inne podzbiory cech do zastosowania w tworzeniu nowych, hybrydowych metod diagnostycznych.