

Gliwice, 20.03.2025

Recenzja rozprawy doktorskiej mgr Jagody Głowackiej-Walas

Informatyczna analiza danych multiomicznych

w oparciu o techniki uczenia maszynowego metody statystyczne

Ukończonej na Wydziale Matematyki i Nauk Informacyjnych
Politechniki Warszawskiej

Pod opieką promotora dr hab. inż. Tomasza Gabina, prof. PW

Tematyka i cel pracy, problem badawczy i jego znaczenie

Przedstawiona mi do recenzji rozprawa doktorska powstała na Wydziale Matematyki i Nauk Informacyjnych Politechniki Warszawskiej pod kierunkiem Profesora Tomasza Gambina. Praca to powstała w formie tak zwanego doktoratu wdrożeniowego i realizowana była przez Doktorantkę zatrudnioną w firmie Transition Technologies-Science (TTSI).

Tematyka rozprawy obejmuje analizę danych multiomicznych za pomocą metod statystycznych oraz technik uczenia maszynowego. Podstawowy cel analiz związany był z budowaniem modeli predykcyjnych umożliwiających przewidywanie wystąpienia napadów padaczkowych lub lekooporności oraz identyfikację biomarkerów procesu chorobowego u dzieci ze stwardnieniem guzowatym, porównanie metod selekcji cech w analizie multiomicznej oraz zaproponowanie techniki pozwalającej na uzyskanie najlepszego zestawu klasyfikatorów w warunkach niekompletności danych.

Rozwój bioinformatyki jest napędzany przez dynamiczny postęp w technologiach wysokoprzepustowych, które znacząco zwiększają możliwości przetwarzania próbek. Skutkuje to generowaniem ogromnych wolumenów danych, wymagających zaawansowania specjalistycznych algorytmów do ich integracji i analizy. Analiza danych multiomicznych odgrywa kluczową rolę w zrozumieniu złożonych procesów biologicznych, ponieważ umożliwia integrację informacji z różnych warstw molekularnych, takich jak genomika, transkryptomika, proteomika i metabolomika. Dzięki temu podejściu możliwe jest uzyskanie szerszego

Politechnika Śląska
Wydział Automatyki, Elektroniki i Informatyki
Katedra Sieci i Systemów Komputerowych
ul. Akademicka 16, pok. 414, 44-100 Gliwice
+48 32 237 21 51 / +48 237 27 33
Aleksandra.gruca@polsl.pl

NIP 631 020 07 36
ING Bank Śląski S.A. o/Gliwice 60 1050 1230 1000
0002 0211 3056

obrazu funkcjonowania organizmu i poznanie kluczowych ścieżek molekularnych związanych z badanymi procesami co pozwala na identyfikację nowych biomarkerów i przez to przyczynia się do rozwoju medycyny spersonalizowanej i precyzyjnej. W przypadku chorób rzadkich, gdzie tradycyjne metody diagnostyczne często zawodzą, analizy multiomiczne stanowią alternatywne narzędzie do odkrywania nowych mechanizmów chorobotwórczych i poprawy diagnostyki, a tym samym poprawy jakości życia pacjentów.

Mimo trwającego już ponad dwie dekady rozwoju metod maszynowego uczenia oraz pojawienia się coraz większej liczby dostępnych danych pochodzących z różnych „omik”, efektywna integracja tych danych pozostaje wyzwaniem. Techniczne i eksperymentalne błędy systematyczne, wprowadzone już na etapie zbierania danych, związane z różnymi protokołami eksperymentalnymi przeprowadzanymi w różnych laboratoriach, wymagają zastosowania odpowiednich technik normalizacyjnych dla heterogenicznych danych. Liczebność próbek, znacznie przekraczająca liczbę badanych pacjentów, oraz związana z tym „klątwa wielowymiarowości”, nieodłączny w analizie danych multiomicznych problem występowania brakujących danych, a także kwestie związane z interpretowalnością wyników zwracanych przez algorytmy powodują, że analiza danych multiomicznych, a w szczególności wykorzystywanie jej w praktyce klinicznej, wciąż pozostaje rzadkością.

W niniejszej rozprawie pani mgr inż. Jagoda Głowcka-Walas podejmuje problematykę niektórych z wymienionych powyżej wyzwań analizy danych multiomicznych. Przedstawione w rozprawie metody koncentrują się na poszukiwaniu biomarkerów przy analizie danych z pojedynczej omiki, opracowanie metod wspomagających wybór algorytmu selekcji cech, wykorzystanie technik uczenia zespołowego do analizy zbiorów danych z brakującymi wartościami oraz stworzeniu narzędzia PlayOmics umożliwiającego budowę modeli predykcyjnych. W kontekście coraz większych możliwości eksperymentalnych pozwalających na analizę różnych omik oraz występującej luki między stosowaniem wyników zintegrowanej analizy danych w zastosowaniach klinicznych, wybrany problem badawczy jest istotny i aktualny oraz bez wątpienia wpisuje się w obecne trendy w dziedzinie bioinformatyki oraz biologii obliczeniowej.

Charakterystyka rozprawy

Przedstawiona praca doktorska składa się ośmiu rozdziałów.

Pierwszy rozdział stanowi wprowadzenie do rozprawy i koncertuje się przedstawieniu cel i zakres pracy, tezy pracy, wkład autorki w powstałe rozwiązania oraz dane analizowane w ramach doktoratu.

Rozdział drugi zatytułowany jest „**Podejścia i wyzwania w analizie multiomicznej**”. W rozdziale tym, stanowiącym wprowadzenie do tematyki pracy, doktorantka przedstawia kluczowe aspekty związane z przetwarzaniem i interpretacją danych multiomicznych, ogólny kontekst analiz multiomicznych oraz opisuje rodzaje danych omicznych wraz z metodami ich przetwarzania. Następnie omawia zastosowania analiz multiomicznych w medycynie i biologii. Kolejne podrozdziały skupiają się na etapach przetwarzania danych, w tym na przygotowaniu i eksploracji danych, redukcji wymiarowości oraz modelowaniu, które obejmuje analizę statystyczną, uczenie maszynowe, ocenę i walidację modeli, a także ich zastosowania w praktyce klinicznej i interpretowalność. Przedstawione są główne wyzwania analizy, takie jak duży rozmiar i złożoność danych, efektywna integracja danych omicznych, mała liczba próbek (szczególnie w kontekście

chorób rzadkich), problem brakujących danych oraz niezbalansowanie klas. Rozdział kończy się omówieniem dobrych praktyk i rekomendacji dla analizy danych multiomicznych, które mają na celu poprawę jakości i efektywności badań.

Rozdział trzeci „**Faza badawczo-rozwojowa**” opisuje wyniki badań wykonanych w ramach projektu EPISTOP oraz EPIMARKER. W projekcie EPIMARKER doktorantka pełniła rolę kierownika projektu ze strony firmy TTSL. W ramach projektu EPISTOP analizowano dane transkryptomiczne, proteomiczne, wybrane dane metabolomiczne, miRNA oraz analizowano DNA. W ramach prac porównano klasyfikatory XGBoost, Lasso oraz regresję logistyczną. Finalnie analizy porównawcze wykonano dla klasyfikatorów w oparciu o regresję logistyczną dla wszystkich możliwych jedno- dwu- i trzelementowych kombinacji zmiennych. Z kolei opisane w doktoracie badania wykonane w ramach projektu EPIMARKER, gdzie analizowano dane kliniczne, proteomiczne, miRNA oraz RNA. Zastosowano tu dwa podejścia: budowa jednego dużego modelu wykorzystującego wszystkie cechy, które pozostały po etapie filtrowania. To podejście przyniosło niezadowolającą jakość predykcji. Budowa wielu małych modeli regresji logistycznej (z dwoma lub trzema zmiennymi wejściowymi). To podejście zostało przyjęte w oparciu o obiecujące wyniki uzyskane przy użyciu małych modeli w projekcie EPISTOP. W obydwu projektach, dużym wyzwaniem była niewielka liczba pacjentów (34 i 65) w obydwu projektach.

W kolejnym rozdziale pt. „**Metody selekcji cech w analizie multiomicznej**” doktorantka koncentruje się na problemie, jakim jest wysoka wymiarowość danych multiomicznych, gdzie liczba cech często znacznie przekracza liczbę próbek ($p \gg n$). W rozdziale tym przedstawiono wyniki dla analiz porównania metod selekcji cech dla pięciu otwartych zbiorów danych wykonanych w ramach pięciu eksperymentów: (1) porównanie wybranych metod selekcji cech (jednowymiarowe: AUC, IG; wielowymiarowe: MRMR; wbudowane: RF, Lasso; oraz DIABLO) w różnych ustawieniach. (2) Analiza pokrywania się oraz typów cech omicznych wybranych przez różne metody, aby zrozumieć ich stabilność i tendencje do preferowania określonych typów danych. (3) Badanie czy uwzględnienie niekompletnych obserwacji w zbiorach danych może poprawić wyniki klasyfikacji, ze szczególnym uwzględnieniem metod jednowymiarowych (AUC, IG), które radzą sobie z brakującymi danymi. (4) Analiza wpływu redukcji liczby próbek na wydajność różnych metod selekcji cech. (5) Porównanie czasu obliczeniowego wymaganego przez różne metody selekcji cech. W podsumowaniu, stwierdzono, że nie istnieje jedna uniwersalna metoda selekcji cech i zawsze należy wybierać metodę selekcji cech w oparciu o charakterystykę zbioru. Równocześnie zostało stworzone narzędzie symulacyjne narzędzia wspomagające proces wyboru właściwej metody selekcji cech.

Rozdział 5 „**Metody uczenia zespołowego adresujące braki danych**” poświęcony jest metodom uczenia zespołowego, które są stosowane w celu rozwiązania problemu braków danych w analizie multiomicznej. W rozdziale analizowano dwa główne podejścia do wyboru zespołu klasyfikatorów: statyczne i dynamiczne. Problem selekcji statycznej polega na wyborze stałego zestawu klasyfikatorów przed fazą predykcji. Doktorantka zaproponowała tu modyfikację powszechnie stosowanej metryki mierzącej niezgodność między klasyfikatorami. Kluczowym pomysłem było traktowanie brakujących predykcji jako formy niezgodności między modelami. W ten sposób proces selekcji ma na celu znalezienie kolejnych klasyfikatorów do zespołu, które mogą dokonywać predykcji dla obserwacji, których poprzednie modele nie były w stanie przetworzyć z powodu brakujących danych. Selekcja dynamiczna polega na wyborze podzbioru klasyfikatorów z puli dla każdej nowej obserwacji na podstawie jej charakterystyki, w tym wzorca brakujących danych.

W szczególności do oceny próbek wykorzystywane są tylko takie modele, które dla danej wykorzystują te dane, które dostępne są dla danej próbki. Eksperymenty w rozdziale piątym przeprowadzono głównie na rozszerzonej wersji zbioru danych BRCA z TCGA, który zawierał dane dotyczące zmienności liczby kopii ($BRCA_R$). Aby symulować scenariusze z większą liczbą brakujących danych, zarówno zbiory walidacyjne, jak i testowe poddano losowemu maskowaniu, zapewniając, że co najmniej 20% wierszy dla każdej omiki zawiera brakujące wartości. Ocena zaproponowanych metod koncentrowała się na liczbie brakujących predykcji w zbiorze testowym oraz dokładności dokonanych predykcji. Wyniki wykazały, że zaproponowane podejście do selekcji statystycznej, szczególnie z wykorzystaniem poprawionej metryki $Miss_{corr}$, było skuteczne w minimalizowaniu brakujących predykcji przy zachowaniu dobrej dokładności klasyfikacji przy mniejszym rozmiarze zespołu klasyfikatorów w porównaniu z poleganiem wyłącznie na jakości predykcji

Przedostatni rozdział rozprawy pt. „**Konstrukcja potoku przetwarzania – playOmics**” zawiera opis pakietu R o nazwie playOmics. Narzędzie to zostało stworzone, aby rozwiązać wyzwania napotykane w analizie danych multiomicznych, wykorzystując doświadczenia z projektów EPISTOP i EPIMARKER. Głównym celem playOmics jest uproszczenie procesy analizy złożonych danych omicznych, umożliwienie budowy skutecznych modeli predykcyjnych, jednocześnie priorytetyzując interpretowalność, łatwość użycia i powtarzalność analiz w celu identyfikacji markerów prognostycznych. Etapy analizy multiomicznej, które umożliwia narzędzie to przygotowanie danych, selekcja cech, trenowanie i walidacja modeli. W odniesieniu do wyników przedstawionych w ramach pracy doktorskiej w narzędziu zaimplantowano metodę tworzenia niewielkich m -czynnиковych modeli regresji logistycznej, narzędzie symulacyjne wspomagające selekcję cech, oraz narzędzie wyboru modeli do analizy danych brakujących przy użyciu metryki $Miss_{corr}$. Działanie pakietu playOmics zaprezentowano dla zbiorów danych BRCA oraz THCA.

Recenzowana praca stanowi doktorat wdrożeniowy, dlatego istotnym elementem oceny jest tu czy przedstawione rozwiązania zostały zaimplantowane w postaci narzędzia i wykorzystane w ramach istniejących, praktycznych projektów. W rozdziale 6, omówionym powyżej, opracowane metody zostały zaimplementowane w pakiecie PlayOmics, w rozdziale 7 „**Implementacja playOmics w projekcie DIPGen**” opisano wykorzystanie narzędzia playOmics jako jednego z modułów systemu analizy danych wykorzystywanego w projekcie DIPG koordynowanym przez Centrum Zdrowia Dziecka w Warszawie.

Ostatni rozdział pracy zwięźle opisuje wnioski z wykonanych analiz oraz kierunki dalszych prac.

Opinia o rozprawie

Należy podkreślić, iż problem, którego podjęła się rozwiązać Doktorantka jest niewątpliwie ważny i wpisujący się w trendy najnowszych badań w dziedzinie. Analiza danych multiomicznych, mimo dużego potencjału i zaawansowanych technologii pomiarowych, wciąż nie jest szeroko stosowana w praktyce klinicznej ze względu na złożoność i heterogeniczność danych, trudności w integracji różnych warstw omicznych oraz problemy z brakującymi danymi. Dodatkowo, analiza ta wymaga dużych zasobów obliczeniowych, a modele predykcyjne często są trudne do interpretacji klinicznej, co ogranicza ich praktyczne zastosowanie. Wreszcie, dla wielu chorób brakuje wystarczającej liczby próbek do wiarygodnej walidacji i generalizacji wyników.

Przedstawiona mi do oceny praca doktorska adresuje niektóre z wyzwań związanych z analizą danych multiomicznych. W rozdziale trzecim autorka mierzy się z praktycznymi problemami analizy danych multiomicznych w chorobach rzadkich, związanymi z częstą w realnych projektach koniecznością analizy niewielkiej liczby próbek pacjentów. Rozdział czwarty opisuje opracowanie narzędzia symulacyjnego, które wspiera użytkowników w wyborze odpowiednich metod selekcji cech dla ich specyficznych zbiorów danych multiomicznych. W rozprawie przedstawiono wpływ różnych metod selekcji cech na wyniki klasyfikacji oraz wskazówki dotyczące ich wyboru w zależności od charakterystyki danych. Rozwiązanie to odpowiada na wyzwania związane z wysoką wymiarowością danych oraz brakiem jednej, uniwersalnie optymalnej metody w analizie multiomicznej. Rozdział piąty przedstawia badania dotyczące zastosowania algorytmów uczenia zespołowego w celu rozwiązania problemu brakujących danych w zbiorach danych multiomicznych. Zaproponowano zmodyfikowaną metrykę różnorodności klasyfikatorów, uwzględniającą szczególnie obserwacje niezaklasyfikowane z powodu brakujących danych. Opisane w tym rozdziale prace stanowią wkład w trwające wyzwanie związane z analizą niekompletnych zbiorów danych multiomicznych. Stworzone w ramach pracy narzędzie playOmics zostało zaprojektowane z silnym naciskiem na zwiększenie interpretowalności analiz oraz zapewnienie reprodukowalności. Zaimplementowanie graficznego interfejsu użytkownika (GUI), szczegółowe rejestrowanie przebiegu eksperymentów oraz generowanie włączenie do pakietu narzędzi pozwalających na interpretowanie wyników modeli (wartości SHAP) bezpośrednio odpowiadają na potrzebę bardziej przejrzystych i zrozumiałych analiz multiomicznych, gdzie reprodukowalność jest kluczowa dla walidacji i dalszego wykorzystania wyników badań. Przy ocenie pracy należy też zwrócić uwagę na praktyczne aspekty przedstawionych badań.

W ostatniej części rozprawa prezentuje zastosowanie narzędzia playOmics (jako elementu aplikacji GeneSense) w projekcie DIPGen do klasyfikacji binarnej pacjentów na podstawie danych multiomicznych. Demonstruje to rzeczywistą przydatność i zastosowanie playOmics w środowisku badań klinicznych, szczególnie w kontekście choroby rzadkiej. Ponadto, metodologie klasyfikacji oparte o budowę niewielkich *m*-czynnиковych modeli regresji logistycznej zaimplementowane w playOmics były wcześniej wykorzystywane w projektach EPISTOP i EPIMARKER, co jest kolejnym argumentem wskazującym, że zaprojektowane przez panią mgr. inż. Jagodę Głowacką-Walas metody wykorzystywane są do poszukiwania nowych biomarkerów w projektach naukowych.

Spis literatury składa się ze 138 pozycji. Jak na pracę doktorską, nie jest to duży przegląd, jednak oddaje on aktualny stan wiedzy w zakresie, którego dotyczy rozprawa. Wiele z cytowanych pozycji to prace najnowsze co stanowi dowód na to, iż Doktorantka orientuje się w aktualnych badaniach prowadzonych w tematyce rozprawy. Świadczy to również o tym, że problem, którego podjęła się rozwiązać oraz wykorzystane w tym celu metody wpisują się w najnowsze trendy badań związanych z analizą i integracją danych multiomicznych.

Pozytywnie też oceniam fakt, iż kody źródłowy metod narzędzia playOmics został udostępniony środowisku naukowemu na platformie GitHub wraz z konfiguracją pozwalającą na uruchomienie narzędzia w kontenerze.

Podsumowując, uważam, iż w swojej pracy Doktorantka zastosowała odpowiednie metody badawcze w celu rozwiązania problemu oceny efektywności różnych metod statystycznych oraz technik uczenia maszynowego w analizie danych multiomicznych w obszarach selekcji cech oraz brakujących danych. Cel rozprawy został osiągnięty. Zastosowane w pracy podejścia do przetwarzania danych wskazują na dobrą znajomość przez Doktorantki nowoczesnych i efektywnych metod stosowanych w dziedzinie biologii obliczeniowej.

Uwagi krytyczne i dyskusyjne

Należy podkreślić, że problem, który podjęła się rozwiązać mgr inż. Jagoda Głowacka-Walas jest niewątpliwie ważny i trudny i na początku tej części chciałbym podkreślić, że nie znalazłam w przedstawionych wynikach żadnych zasadniczych błędów merytorycznych. Wszystkie poniższe uwagi wynikają z chęci podjęcia dyskusji i dialogu na temat niektórych aspektów pracy. Uwagi te nie obniżają mojej pozytywnej oceny pracy.

Za istotny mankament pracy uważam w części teoretycznej raczej dość słaby przegląd literatury zarówno dla metod selekcji cech jak i metod radzenia sobie z brakującymi danymi w analizie danych multiomicznych, a także brak szerszego porównania zaproponowanych rozwiązań z metodami *state-of-the-art*. Przynajmniej w przypadku prac związanych z omówieniem metod selekcji cech Doktorantka jako jedną z metod do porównania wybrała metodę DIABLO, która dedykowana jest do analizy danych multiomicznych. Natomiast w rozdziale związanym z podejściem do problemu brakujących wartości, zaproponowana metoda nie jest porównywana do żadnych innych znanych z literatury metod radzenia sobie z brakującymi danymi, a w pracy nie podjęto nawet dyskusji teoretycznej porównującej zaproponowane podejście z istniejącymi metodami. Istnieje wiele prac przeglądowych, np. praca przeglądowa Flores et al. [10.3389/frai.2023.1098308], czy praca Hornung et. al. [doi.org/10.1002/wics.1626], obie z nich cytowane w rozprawie, które omawiają szereg metod wykorzystywanych do radzenia sobie z problemem brakujących danych.

W przypadku wielokrotnego budowania modeli istnieje ryzyko, że jeden z kilku tysięcy testowanych modeli przypadkowo dopasuje się do dodanych dając nadmiernie optymistycznie wyniki. W pracach opisanych w rozdziale trzecim zastosowano odpowiednie metody mitygacji takiej sytuacji przez zastosowanie symulacji Monte Carlo oraz wykorzystywanie oddzielnego zbioru walidacyjnego. Czy stworzone w ramach doktoratu narzędzie PlayOmics zawiera mechanizmy wymuszające na użytkowniku stosowanie metod przeciwdziałania przeuczeniu modeli?

W pracy podkreślono ważność interpretacji wyników zwracanych przez modele. W pracy możliwość interpretowalności została zrealizowana z jednej strony poprzez zaproponowanie interpretowalnych modeli regresji, a z drugiej przez zaimplementowanie metody SHAP w narzędziu PlayOmics. Czy w analizach przeprowadzonych i przedstawionych w ramach pracy doktorskiej dla zbiorów BRCA i THCA uzyskane wyniki pokrywały się? Jeśli nie, co mogło być przyczyną różnic?

Uwagi redakcyjne:

W podrozdziale 4.6 warto by było umieścić rekomendacje w zbiorczej tabeli, ułatwiłoby to podsumowanie i syntezę wniosków uzyskanych w ramach badań przeprowadzanych w rozdziale 4.

Dobrze byłoby zadbać o bardziej staranne przygotowanie rysunków do pracy. Część z pokazanych na nich wyników jest trudno zinterpretować (np. rys 4.5), dla niektórych „ucięto” opisy osi (np. rys 6.4 D, rys 6.7), a część informacji jest nieczytelna (rys 6.4 C i F)

Podsumowanie

Pani mgr inż. Jagoda Głowacka-Walas przedstawiła rozprawę doktorską rozwiązującą aktualny problem naukowy, która przyczyni się do rozwoju reprezentowanej dyscypliny naukowej. Rozprawa zawiera oryginalne rozwiązanie problemu naukowego, a Doktorantka wykazała, że zarówno posiada ogólną wiedzę teoretyczną w dyscyplinie informatyka techniczna i telekomunikacja jak i umiejętność prowadzenia pracy naukowej.

Pani mgr inż. Jagoda Głowacka-Walas jest współautorką 9 publikacji naukowych, w tym w trzech z nich występuje 3 jako członek konsorcjum EPISTOP. Jedna z tych publikacji została opublikowana w prestiżowym czasopiśmie Nature Communications, gdzie pani mgr inż. Jagoda Głowacka-Walas posiada pierwsze autorstwo współdzielone. Jest to niewątpliwie duże osiągnięcie. Wg bazy danych Scopus większość z czasopism, w których publikuje doktorantka to czasopisma z dziedziny medycyny, neurologii, pediatrii co wskazuje na aplikacyjny charakter przeprowadzonych prac i ich silne powiązanie z projektem. Ponadto wyniki prac zostały zaimplementowane w postaci narzędzia i wykorzystane do analizy danych w projekcie naukowym. W doktoratach wdrożeniowych, biorąc pod uwagę charakter prac, które koncertują się bardziej na aspekcie wdrożenia produktu, zwyczajowe wymagania publikacyjne są zwykle niższe niż w przypadku doktoratów tradycyjnych. Dlatego też bardzo wysoko oceniam fakt, iż Doktorantka podjęła wysiłek publikacyjny zakończony sukcesem w tak dobrych czasopismach.

Biorąc pod uwagę ocenę przedstawioną w niniejszej recenzji, stwierdzam, że przedstawiona do oceny praca doktorska w pełni odpowiada warunkom określonym w Art. 187 ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (tekst jednolity Dz. U. z 2024 r. poz. 1571 z późn. zm.) i na tej podstawie wnoszę do Wysockiej Rady Naukowej Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Warszawskiej o dopuszczenie mgr inż. Jagody Głowackiej -Walas do dalszych etapów przewodu doktorskiego.

dr inż. hab. Aleksandra Gruca, prof. PŚ.