UNIWERSYTET WARSZAWSKI

Instytut Informatyki
Uniwersytet Warszawski
ul. Banacha 2
02–097 Warszawa
POLSKA

dr hab. Bartosz Wilczyński
profesor uczelni
Phone: +(48 22) 5544 577
Fax: +(48 22) 5544 400
e-mail: bartek@mimuw.edu.pl

Warszawa, 4. maja 2025 r.

Recenzja rozprawy doktorskiej pt. „Informatyczna analiza danych multiomicznych w oparciu o techniki uczenia maszynowego i metody statystyczne" przedstawionej przez mgr inż. Jagodę Głowacką – Walas

Recenzja niniejsza została sporządzona na zlecenie Rady Naukowej Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Warszawskiej zgodnie z wymogami ustawy dotyczącej procedur nadawania stopnia doktora. Recenzja ta składa się najpierw ze skrótowego opisu merytorycznej zawartości rozprawy, następnie zawiera moje uwagi krytyczne dotyczące cech rozprawy, które stanowią pewne usterki czy niedociągnięcia i zakończona jest podsumowaniem.

Opis rozprawy

Praca złożona przez p. Głowacką-Walas jest obszerna, zawiera wyniki opublikowane już wcześniej w kilku publikacjach. Pomijając rozdział 1, stanowiący wprowadzenie do rozprawy, rozprawa składa się z kilku głównych części (jak opisuje to sama autorka):

- Wstęp metodologiczny w rozdziale 2.
- Wyniki prac w projektach EPISTOP i EPIMARKER w rozdziale 3.
- Wyniki eksperymentów dotyczących metod selekcji cech i uczenia zespołowego w rozdziałach 4-5
- Model narzędzia PlayOmics w rozdziale 6 i przykładowa implementacja potoku w rozdziale 7.

Z punktu widzenia wkładu autorki do badań naukowych, niewątpliwie jej udział w konsorcjum EPISTOP i projektach związanych z tym dużym przedsięwzięciem naukowym zaowocowały najbardziej widocznymi publikacjami. Z drugiej strony, niestety te wyniki budzą największe wątpliwości dotyczące metodologii. Z kolei wyniki przedstawione w późniejszych rozdziałach, czy to wyniki symulacji i eksperymentów z rozdziałów 4-5, czy implementacja narzędzia playOmics w rozdziałach 6-7, cechują się dużo lepszym warsztatem obliczeniowym, choć niestety nie mają takiego wpływu na środowisko naukowe.

Z mojego punktu widzenia, wszystko co w rozprawie jest napisane, wskazuje, że autorka miała styczność z dużym projektem i rzeczywistymi danymi klinicznymi, co niestety - mimo niewątpliwie dużego wysiłku autorki włożonego w konstrukcję efektywnych klasyfikatorów, nie doprowadziło do budowy modeli przełomowych dla predykcji lekooporności czy stopnia dolegliwości ataków w badanych przypadkach. Niemniej, doświadczenie zdobyte przez autorkę w tych projektach zaowocowało wykonaniem przez nią kolejnych eksperymentów (opisanych w rozdziałach 4-5) i przygotowanie narzędzia Playomics, które może posłużyć kolejnym badaczom. Trudno jednoznacznie ocenić, czy możliwe byłoby uzyskanie lepszych wyników klasyfikacji w projekcie EPISTOP, jednak biorąc pod uwagę skąpość dostępnych danych i trudność niektórych z problemów, wydaje się to bardzo prawdopodobne, że dla tych danych nie byłoby to możliwe.

Uwagi krytyczne

Istotna wątpliwość dotycząca wyboru modeli predykcyjnych na podstawie testów permutacyjnych

Autorka przedstawia rozwiązanie korzystające z "małych" modeli regresji logistycznej w projekcie EPISTOP. Wykorzystuje tam "małe" modele regresji logistycznej i ocenia je na podstawie testów permutacjyjnych. W szczególności na rysunku 3.6, jeśli dobrze rozumiem, przedstawia rozkład miary MCC na zbiorach testowych dla bardzo wielu modeli i zwraca uwagę na dominującą wartość 0, oznaczającą niską jakość predykcji. Później wykonuje testy na permutowanych danych i w ten sposób wybiera próg, aby co najwyżej 5 procent klasyfikatorów losowych było lepszych niż te wybrane ostatecznie. Nie jestem przekonany, czy to jest wystarczająca metoda. W szczególnośći, gdyby okazało się, że rozkład testowych wartości MCC dla permutowanych danych jest podobny do tego na wykresie 3.6, należałoby uznać, że zmienne objaśniające nie wnoszą istotnej informacji nt. zmiennej objaśnianej i nie można byłoby wybierać po prostu 5 procent najlepszych losowych klasyfikatorów. Fakt, że autorka spośród 22100 modeli, które mają być lepsze niż 0.05 losowych wybiera 100 modeli, sugeruje, że w istocie tych modeli jest około 0.05 co sugerowałoby, że mamy do czynienia z niemalże losowymi klasyfikatorami.

Chciałbym, aby autorka jakoś odniosła się do tej wątpliwości. Tzn. uważam, że nawet jeżeli w ostatecznym rozrachunku wynik tych badań jest negatywny (tzn. żaden z tych modeli nie jest lepszy niż to czego spodziewamy się w modelu zupełnie losowym, bez

związku pomiędzy zmiennymi objaśniającymi a objaśnianymi), to nie przekreśla całości dorobku doktorantki, ale jednak chciałbym wiedzieć, czy tam naprawdę jest jakaś nielosowa zależność, czy jednak ostatecznie zwrócone modele w zasadzie nie różnią się od tego, czego moglibyśmy się spodziewać, gdybyśmy takie dane wygenerowali zupełnie losowo, bez związku pomiędzy zmienną objaśnianą i predyktorami.

Drobniejsze uwagi natury redakcyjnej i stylistycznej

W rozdziale 2, autorka opisując działy takie jak genomika, transkryptomika i proteomika, czyni to w sposób niezmiernie skrótowy. Uwzględnia tu też wyłącznie typy danych, które później sama analizuje (pomijając np. dane takie jak ChIP-Seq czy Hi-C). Ta część mogłaby zyskać na bardziej przekrojowym potraktowaniu materiału.

W rozdziale 3, rysunek 3.3 przedstawia krzywe ROC dla 3 modeli: z 1 zmienną, wszystkimi zmiennymi i "optymalny" z 6 zmiennymi. Niestety, krzywa ROC dla "optymalnego"modelu jest wyraźnie gorsza niż dla modelu z 1 zmienną. Autorka nie komentuje tego faktu, choć powinna go zauważyć. Rozumiem, że model wybrany poprzez optymalizację MCC może się nie sprawdzać, ale jednak wymagałoby to jakiegoś komentarza.

Autorka często używa sformułowania "adresować" w znaczeniu będącym kalką angielskiego słowa "address", czyli "znaleźć rozwiązanie" czy "odpowiedzieć na wątpliwość", co nie jest zgodne z polskim znaczeniem tego słowa, ale pojawia się coraz częściej w potocznej polszczyźnie.

Podsumowanie

Podsumowując, praca mgr inż. Jagody Głowackiej-Walas podejmuje istotne problemy w dziedzinie bioinformatyki. Mimo pewnych niedociągnięć i usterek, o których pisałem wcześniej, praca stanowi niewątpliwie wkład do dyscypliny informatyka w dziedzinie nauk technicznych i świadczy o tym, że autorka uzyskała poziom wiedzy i samodzielności naukowej oczekiwany na etapie doktoratu. W związku z tym uważam, że **rozprawa doktorska spełnia ustawowe wymagania wobec prac doktorskich** i może zostać skierowana do kolejnych etapów przewodu doktorskiego.

Z poważaniem,

Bartosz Wilczyński