

Recenzja rozprawy doktorskiej

Autor: mgr Piotr Radziński

Tytuł: Computational methods for leading-edge mass spectrometry techniques

Metody obliczeniowe dla wiodących technik spektrometrii mas

Promotor: prof. dr hab. Anna Gambin

Ogólna charakterystyka rozprawy

Przedłożona do recenzji rozprawa doktorska jest napisana w języku angielskim, liczy 89 stron tekstu, składa się z sześciu rozdziałów tekstu oraz wykazu literatury. Praca jest poświęcona opisom różnych metodologii analizy danych otrzymywanych technikami eksperymentalnymi spektrometrii mas obejmujących trzy z czterech jej rozdziałów analitycznych oraz opisom przykładów konstruowania potoków analiz biostatystycznych w badaniach medycznych. Tematyka pracy jest bardzo współczesna, odnosi się do najnowocześniejszych technik eksperymentalnych proteomiki i genomiki. Ma ponadto charakter i walor interdyscyplinarny, proponowane i rozwijane algorytmy przedstawiane są w kontekście opisów instalacji laboratoryjnych używanych do uzyskania danych, a także w kontekście zastosowań biologicznych, chemicznych czy medycznych, wyników uzyskanych przy ich użyciu.

Rozdział pierwszy pracy, wstępny, obejmuje ogólne opisy problemów stojących za konstruowanymi algorytmami oraz idei ich rozwiązania. Pierwszym z rozwiniętych w ramach pracy podejść jest algorytm wyznaczania masy monoizotopowej polimerów białek/peptydów oraz oligonuklotydów w wysokorozdzielczych spektrach masowych, zawierających obwiednie izotopowe. Problematyka ta jest dalej dokładniej opisana w rozdziale 2. Drugim podejściem zbudowanym w pracy Doktoranta jest metodologia kompresji danych obrazowania medycznego z pikselami definiowanymi przez widma masowe, bazująca na stworzeniu i wytrenowaniu głębokiej sieci neuronowej typu enkoder – dekoder. Opracowana

metodologia jest następnie opisana dokładniej w rozdziale 3. Kolejny, trzeci algorytm zaproponowany w recenzowanej pracy jest związany z techniką eksperymentalną spektrometrii masowej sprzężonej z wzbudzaną indukcyjnie plazmą. Zaproponowany (rozwinięty) algorytm pozwala na bardzo precyzyjną ocenę stosunku stężeń różnych izotopów tego samego pierwiastka w badanych próbkach. Ta problematyka jest dalej rozwinięta i dokładniej opisana w rozdziale 4. Wreszcie ostatnim analitycznym zagadnieniem omawianym w pracy jest analiza biostatystyczna kohort w studiach medycznych. W tym zakresie biostatystyczne analizy danych uzyskanych przez doktoranta w ramach współpracy z grupami medyków i klinicystów. Dokładny opis konstrukcji potoków analiz jest przedstawiony w rozdziale 5.

Rozdział drugi poświęcony jest problematyce wyznaczania masy monoizotopowej cząstek białek/peptydów oraz oligonukleotydów w wysokorozdzielczych spektrach masowych, zawierających obwiednie izotopowe. Na początku tego rozdziału Doktorant przedstawia motywację do rozwijania metod wyznaczania masy monoizotopowej białek/peptydów lub oligonukleotydów na podstawie widm masowych wysokiej rozdzielczości. Znajomość (ocena) masy monoizotopowej może być użyta do identyfikacji badanych związków / polimerów. Doktorant przedstawia przegląd istniejących algorytmów wyliczania masy monoizotopowej cząstki na podstawie widma obwiedni izotopowej. Jednym z nowszych metod jest algorytm MIND opracowany w grupie naukowej, z którą związany jest także Doktorant. Opisuje pewne wady dotychczasowych algorytmów. Następnie w punkcie 2.1 przedstawia nowy algorytm Envemind, który ulepszeniem i rozwinięciem algorytmu MIND. Nowy algorytm Envemind działa dwuetapowo. Dwa etapy algorytmu przedstawione są w postaci schematu blokowego na rysunku 2.1. W pierwszym etapie, na podstawie bazy danych Uniprot oraz istniejącego algorytmu Isospec do symulacji struktur białkowych kalibruje się dwa modele liniowe (2.1) oraz (2.5) których parametry są, w drugim etapie algorytmu wykorzystywane do dopasowania modelu cząstki do obserwowanego widma obwiedni izotopowej. W drugim etapie algorytmu Envemind wykorzystuje się sformułowaną w literaturze w 1995 roku, kluczową koncepcję „aweraginy”, wirtualnej (hipotetycznej) cząstki zbudowanej z atomów C, H, N, O, S, w proporcja odpowiadających obserwowanym proporcjom tych atomów, oraz ich izotopów, w znanych białkach. Pomysł / koncepcja aweraginy jest powszechnie znana w literaturze i bardzo użyteczna. Doktorant rozwija ideę aweraginy. Interpretuje ją jako związaną / wynikającą z pierwszej składowej głównej tablicy zbudowanej przez symulacje wszystkich cząstek białkowych z bazy Uniprot. Wprowadza drugą wirtualną cząstkę Beweraginę, związaną z drugą składową główną tej tablicy. Dokonuje wyczerpujących porównań algorytmu Envemind z wcześniejszym algorytmem MIND i wykazuje przewagę tego pierwszego w sytuacji, gdy dane eksperymentalne charakteryzują się dużą rozdzielczością i dużą złożonością.

W punkcie 2.1 Doktorant opisuje jeszcze następny algorytm MIND4OLIGOS, który jest wersją algorytmu Envemind dostosowaną do obliczania masy monizotopowej oligonukleotydów, na podstawie ich widm obwiedni izotopowych. Algorytm MIND4OLIGOS

jest uproszczeniem algorytmu Envemind. Wykorzystuje ideę sumowania związków chemicznych, tak jak poprzednio. Jego podstawową ideą jest ocena najbardziej prawdopodobnej odległości między masą monoizotopową a masą szczytową (offsetem) wyrażoną liczbą całkowitą. W podpunkcie 2.2.3 Doktorant przedstawia ocenę jakości algorytmu MIND4OLIGOS, wykorzystując do tego symulacyjny zbiór danych 10000 oligonukleotydów. Wykazuje dobrą jakość zaproponowanego podejścia

Rozdział trzeci pracy poświęcony jest kompresji danych obrazowania medycznego z pikselami definiowanymi przez widma masowe, przy użyciu głębokiej sieci neuronowej typu enkoder – dekoder. Schemat blokowy zaproponowanego algorytmu jest przedstawiony na rysunku 3.1. Do nauczania / nastrojenia sieci wykorzystuje się ideę uczenia kontrastowego, z funkcją kary / straty zapisaną wzorami (3.1), (3.2). Dodatkowo, aby uzyskać pożądane własności sieci definiuje się dwie funkcje straty (3.3) i (3.4). Tak zdefiniowane funkcje straty zapewniają pożądane własności rozkładów cech wyliczanych przez sieć. Aby nauczona sieć enkodera – dekodera posiadała dobre własności odtwarzania zakodowanych obrazów wykorzystuje się jeszcze jedną funkcję porównującą zakodowany i oryginalny obraz, wyliczającą wskaźnik ich średniego, kwadratowego odchylenia, dany wzorem (3.5). Ocenę jakości zbudowanej sieci enkodera – dekodera realizuje się przez wykonanie segmentacji zakodowanych (skompresowanych) i oryginalnych obrazów, z zastosowaniem algorytmu k-średnich i porównaniu uzyskanych segmentów z „prawdziwymi segmentami” uzyskanymi przez zastosowanie odpowiedniej specjalnej procedury (łączącej widma proteomiczne oraz obrazy wybarwione techniką HE badanych próbek). Dla danych dotyczących obrazowania widm proteomicznych moczowego myszy wykazano, w tabeli 3.1, że segmentacja obrazów, metodą k – średnich, daje lepsze wyniki dla obrazów zakodowanych niż dla obrazów oryginalnych. Drugim, większym zbiorem danych dla weryfikacji zaproponowanego algorytmu były obrazy zbudowane z widm masowych pochodzące z biopsji przełyku Barretta. W tabeli 3.2 przedstawiono stopnie kompresji oraz dokładność segmentacji dla tego zbioru danych. Ponownie wykazuje się dobre własności proponowanej metody.

Rozdział czwarty poświęcony jest problematyce bardzo precyzyjnej oceny stosunku stężeń różnych izotopów tego samego pierwiastka w badanych próbkach, na bazie techniki spektrometrii masowej sprzężonej z wzbudzaną indukcyjnie plazmą. Doktorant przedstawia kilka ważnych zastosowań tej metody pomiarowej, w chemii, kosmologii, badaniach środowiskowych. Charakteryzuje specyfikę problemu oceny stosunku stężeń izotopów, związanej z koniecznością porównywania uzyskanych wyników pomiarowych z odpowiednimi wzorcami. Przedstawia trzy sposoby zastosowania wzorców do uzyskania ocen stosunków stężeń izotopów i związane z nimi modele matematyczne. Pierwszy sposób nazywa się metodą wzorca wewnętrznego i jego model opisany jest równaniami (4.1), (4.2). Druga metoda, opisana przez formuły (4.3), (4.4) nazywana jest modelem optymalizowanej regresji. Wreszcie trzecia metoda, „nawiasowania” (bracketing) obejmuje sytuację, gdy pomiar dla wzorca wykonuje się dwukrotnie (przed i po pomiarze dla próbki), jest modelowana równaniem (4.5). W odróżnieniu od dwóch poprzednich metod, w tym

podejściu ocenia się nie samą wartość stosunku stężeń lecz jej odchylenie od wartość stosunku stężeń dla wzorca. Rozwija się metodę oceny ewolucji wartości błędu pomiaru stosunku stężeń przy wielokrotnym powtarzaniu eksperymentu. Przy zmiennej wariancji pojedynczego błędu stosuje się statystykę Q dla zamodelowania błędu pomiarowego. W tabeli 4.2 przedstawia się zbiorcze wyniki zastosowania algorytmu bazującego na modelu (4.5) dla wielu zbiorów danych pomiarowych spektrometrii masowej sprzężonej z wzbudzaną indukcyjnie plazmą.

Rozdział piąty przedstawia opisy kilku projektów badań medycznych / klinicznych, w których Doktorant był zaangażowany w ich analizę statystyczną. Dla każdego z tych projektów zamieszczona jest związana z nim wieloautorska publikacja.

Rozdział szósty stanowi podsumowanie pracy i charakteryzuje możliwości i kierunki dalszych badań.

Ocena rozprawy

Recenzowana rozprawa jest oryginalnym dziełem naukowym. Zawiera wiele nowych, interesujących koncepcji w obszarze analizy danych eksperymentalnych pochodzących z masowych widm proteomicznych uzyskanych najnowszymi technikami.

Praca jest napisana ładnie i jasno. Doktorant wykazuje bardzo wysokie kompetencje w obszarze budowania złożonych modeli matematycznych. Praca dokumentuje umiejętność Doktoranta posługiwania się zaawansowanym warsztatem matematycznym. Opracowanie stabilnych implementacji tych modeli, ich przetestowanie i udostępnienie dowodzi dysponowania przez Doktoranta odpowiednim warsztatem badawczym w zakresie technik informatycznych.

Warsztat naukowy Doktoranta jest szeroki i różnorodny. Doktorant potrafi dopasować metody i modele matematyczne do różnego typu danych. Wykazuje się także (rozdział 5) umiejętnością budowania potoków analiz biostatystycznych zbiorów danych klinicznych.

Praca ma walor interdyscyplinarny. Doktorant przedstawia zasady i charakterystyki technik pomiarowych uzyskiwania widm masowych i potrafi powiązać i dopasować metody modelowania oraz algorytmy obliczeniowe do specyfiki przeprowadzonego eksperymentu.

Spośród wyników naukowych przedstawionych w pracy za najciekawsze i najbardziej wartościowe uważam wyniki przedstawione w rozdziale drugim, dotyczące oceny masy monoizotopowej cząstek białek i oligonukleotydów. Konstrukcje algorytmów obliczeniowych bazują na nowych, interesujących pomysłach. Są opracowane w nawiązaniu i jako rozwinięcie już istniejących metod, które wykazały użyteczność i są często wykorzystywane.

Wreszcie opracowane metody zostały zweryfikowane na bazie dużych zbiorów danych, pochodzących z bazy Uniprot.

Także inne opisywane wyniki naukowe, kompresja danych obrazowania MSI, analizy danych spektrometrii masowej sprzężonej z wzbudzaną indukcyjnie plazmą są bardzo wartościowe.

Publikacje naukowe, których współautorem jest Doktorant wymienione na końcu rozdziału wstępnego, tematycznie związane z doktoratem ukazały się w renomowanych, wysoko punktowanych czasopismach naukowych.

Uwagi krytyczne i dyskusyjne, zauważone błędy, pytania do doktoratu

Wzór (1.1) na str. 10 jest niepoprawny. Brakuje znaków silni w mianowniku.

Wyniki przedstawione w rozdziale 2 są bardzo ciekawe, jednak część z nich jest trudna do prześledzenia i weryfikacji (np. wzór (2.3)) od strony matematycznej. Przydałoby się trochę lepsze ich przedstawienie i umotywowanie.

Zamieszczona w rozdziale 3 konstrukcja sieci enkoder – dekodek jest opisana dobrze i wyczerpująco. Jednak brakuje zbadania jej zastosowania do szerszych zbiorów danych. Rozdział 3 zawiera tylko dwa przykłady możliwości jej zastosowania.

Temat kompresji danych MSI jest bardzo ważny. Byłoby bardzo interesujące, gdyby porównać metodę opracowaną w rozdziale 3 do innych algorytmów literaturowych.

W rozdziale 4 opisane są trzy metody oceny stosunku stężeń izotopów, wzorca wewnętrznego, optymalizowanej regresji, „nawiasowania”. Jednak analizuje się tylko jedną. Warto byłoby dokonać porównania wszystkich trzech metod, zwłaszcza, że Doktorant dysponuje dużym zbiorem danych.

Uwaga na końcu punktu 6.1 o zastosowaniu ChatGPT w pisaniu pracy budzi zastrzeżenia. Nie mam wątpliwości, że recenzowana rozprawa jest oryginalnym opracowaniem naukowym. Jednak jeśli się stosuje metody llm (large language models) to charakter ich zastosowania powinien być bardziej precyzyjnie opisany. Jest możliwe i powszechnie akceptowane używanie narzędzi internetowych oraz ewentualnie llm do korekty błędów edycyjnych oraz weryfikacji poprawności struktury gramatycznej zdań. Jednak logika tekstu, kompozycja, jasność, klarowność są atrybutami autorskimi i nie mogą być cedowane do narzędzi typu ChatGPT.

Pytania:

1. Czy model „lepkiej rolki” (sticky roller) z rys. 2.2 ma jakieś odniesienie w literaturze?
2. Czy Doktorant mógłby dokładniej przedyskutować złożoność obliczeniową problemu minimalizacji wariancji (2.4)? Dlaczego uzyskanie ζ^* wymaga tak długiego czasu? Czy nie dałoby się jakoś usprawnić tych obliczeń?
3. Czy wyniki segmentacji (rys. 3.5) dają się odnieść do histologii pęcherza moczowego myszy? Czy da się nazwać wyodrębnione obszary, np. nabłonek, śluzówka, mięśnie?
4. Jak dokładnie wyglądała konstrukcja algorytmu Monte Carlo opisanego w punkcie 4.3.3? Czy symulacje stochastyczne powtarzano wiele razy dla oceny zmienności?

Konkluzja

Praca stanowi podsumowanie oryginalnych, interdyscyplinarnych badań naukowych, w których formułowano i weryfikowano oryginalne hipotezy badawcze. Dokumentuje wiedzę i kompetencje Doktoranta.

Uwagi krytyczne i dyskusyjne nie mają wpływu na pozytywną ocenę pracy.

Osiągnięcia i oryginalne elementy pracy są na pewno wystarczające do jej ogólnej bardzo dobrej oceny. Stwierdzam, że rozprawa spełnia warunki stawiane pracom doktorskim i wnioskuję o jej dopuszczenie do publicznej obrony.

