prof. dr hab. inż. Marta Kasprzak
Instytut Informatyki
Politechnika Poznańska

Poznań, 17.04.2019

Recenzja rozprawy doktorskiej mgr. inż. Krzysztofa Gogolewskiego pt. „Matrix methods in transcriptomic and metabolomic data analysis"

1. Problematyka naukowa rozprawy

Funkcjonowanie organizmów w dużej mierze opiera się na informacji zakodowanej w cząsteczkach kwasów nukleinowych DNA i RNA obecnych w ich komórkach. Odczytanie tej informacji było od zawsze jednym z podstawowych celów badań z zakresu biologii molekularnej. Początkowo realizowane niezautomatyzowanymi metodami laboratoryjnymi w bardzo małej skali, weszło na wyższy poziom dzięki wynalezieniu technik wysokoprzepustowego pozyskiwania informacji genetycznej. Obecnie do odczytania informacji o sekwencjach DNA i RNA wykorzystuje się m.in. metody mikromacierzowe i sekwenatory nowej generacji. W efekcie jeden taki eksperyment dostarcza dane w ilości wymagającej użycia do ich analizy wyspecjalizowanych programów komputerowych.

Problemy badawcze sformułowane w rozprawie odwołują się do danych pozyskanych z użyciem technik wysokoprzepustowych, które wymagały nowych sposobów analizy. Założenie, że dane na wejściu pochodzą z komórek co prawda homogenicznych, jednak pełniących chwilowo różną rolę w procesach zachodzących w organizmie, co może mieć wpływ na ich interpretację, pociągnęło za sobą dostosowanie używanych do tej pory podejść do tej nowej sytuacji. Temu zadaniu poświęcił się Doktorant, a opracowane przez niego algorytmy wymagały zaangażowania zaawansowanego aparatu matematycznego. Badania zostały poprowadzone dla danych transkryptomicznych, czyli informacji bezpośrednio zawartej w kwasach nukleinowych, ale także metabolomicznych, czyli informacji powiązanej z reakcjami ich produktów. Weryfikacja praktyczna nowych propozycji przeprowadzona została na danych rzeczywistych pozyskanych w instytutach medycznych w eksperymentach z użyciem technik wysokoprzepustowych.

Tematyka rozprawy, choć osadzona w realiach biologicznych, skoncentrowana jest na obliczeniowych aspektach poruszanych zagadnień. Opracowane metody służą dokładniejszej ze statystycznego punktu widzenia analizie danych eksperymentalnych oraz wykrywaniu nowych wzorców w danych, nieosiągalnych wcześniejszymi podejściami. Tematyka ta może więc być przedmiotem rozprawy doktorskiej z dziedziny informatyki.

2. Opinia o rozprawie

Rozprawa składa się z sześciu rozdziałów, z czego cztery zawierają opis rezultatów badań Doktoranta. Pierwszy rozdział wprowadza czytelnika w tematykę rozprawy i krótko streszcza uzyskane wyniki. Więcej o poszczególnych zagadnieniach badawczych i wcześniejszych podejściach czytelnik dowiaduje się ze wstępów do kolejnych rozdziałów dedykowanych rozwiązywanym problemom.

W rozdziale 2 opisano, w jaki sposób algorytm nieujemnej faktoryzacji macierzy może zostać zastosowany do dekompozycji sygnału pochodzącego z eksperymentów na transkryptomie. Wiedza o funkcjonowaniu komórek może być wywiedziona m.in. z pomiaru ekspresji genów, w tym przypadku dokonanego z użyciem technologii mikromacierzy. Dane te zostały pozyskane eksperymentalnie z wielu komórek dających nieidentyczną odpowiedź pomimo tego, że komórki były tego samego rodzaju. Różnica w odpowiedzi wynikać może z uczestnictwa komórek chwilowo w innych procesach w organizmie. Rozbicie sygnału na części składowe umożliwia uzyskanie dokładniejszego obrazu procesów zachodzących w tym fragmencie systemu biologicznego. Doktorant zaproponował nowy schemat przetwarzania z wykorzystaniem wcześniejszych algorytmów opracowanych dla danych z komórek heterogenicznych, którego istotną częścią jest wstępne przygotowanie danych, zwłaszcza selekcja najbardziej znaczącej informacji. Rezultatem jest nowa metoda MPH (*Molecular Processes Heterogeneity*), która sprawdzona została z sukcesem na danych eksperymentalnych pochodzących z komórek nowotworowych neuroblastomy.

Podobnego przeznaczenia algorytm jest przedmiotem opisu w rozdziale 3, przy czym zamiast danych pochodzących z eksperymentów mikromacierzowych mamy do czynienia z wynikami sekwencjonowania nowej generacji. Tym razem zaproponowano inne procedury matematyczne, dostosowane do odmiennego charakteru danych. Faktoryzacja macierzy zastąpiona została rozbiciem macierzy na komponenty składowe, w efekcie prowadzącym do oszczędnej reprezentacji danych. Opracowany został algorytm tRPCA (*truncated Robust Principal Component Analysis*), który realizuje obliczenia w krótszym czasie niż wcześniejsza metoda RPCA, podczas gdy precyzja odpowiedzi pozostaje zachowana. Dodatkowo, w celu zwiększenia skuteczności metody dekompozycji, wprowadzony został do algorytmu nowy komponent w postaci macierzy szumu, która wyrównuje nieregularny niekiedy sygnał biologiczny. Komponenty wynikowe poddawane są klastrowaniu, które ma zobrazować, jakie podtypy komórek można wyróżnić w danych. Walidacja metody przeprowadzona została na danych pochodzących z komórek krwi zdrowego dawcy, w rezultacie wykazano, że uzyskana redukcja danych zachowuje ich reprezentatywność dla całego zestawu, a odfiltrowanie szumu umożliwia wykrycie wzorców w danych.

Uzupełnienie danych nt. ekspresji genów o dodatkową informację pozyskaną za pomocą innych eksperymentów daje szerszą wiedzę o procesach zachodzących w organizmie, stawia jednak większe wyzwanie algorytmom analizy. Taki przypadek opisany został w rozdziale 4 rozprawy, gdzie dodatkowym źródłem informacji są sieci opisujące procesy metaboliczne. Podobnie jak wcześniej, celem jest rozpoznanie podtypów komórek o charakterystycznych wzorcach ekspresji genów odpowiadających procesom, w których biorą udział. Pierwszym etapem metody jest zamodelowanie z użyciem sieci Petriego tego fragmentu sieci

metabolicznej organizmu, który obejmuje enzymy syntetyzowane na bazie genów będących przedmiotem zainteresowania. Tranzycje w sieci są opisane regułami, a do reguł są podstawiane wartości zmiennych wywiedzione z informacji o ekspresji genów: dany gen występujący w regule przyjmuje wartość 1 lub 0 w zależności od tego, czy jest aktywny, czy nie. W ten sposób część sieci jest wyciszana i w takiej postaci poddawana analizie. Zaprojektowana przez Doktoranta procedura przetwarzania ma ułatwić ekstrakcję wzorców z danych i wysnucie biologicznie poprawnych wniosków, co zostało zweryfikowane w praktyce na danych uzyskanych dla komórek nowotworu nerki. Zwrócona została przy tym uwaga na to, aby nadreprezentowane reguły nie wpływały na wnioskowanie. Kod źródłowy realizujący to przetwarzanie został udostępniony do pobrania na GitHubie.

Rozdział 5 zawiera opis badań poszczególnych przypadków danych klinicznych. Doktorant zastosował do ich przetworzenia i wydobycia istotnej biologicznie informacji znane z literatury metody statystycznej analizy danych oraz udostępnione pakiety oprogramowania. Rozprawę zakończył podsumowaniem w rozdziale 6.

Formułując swoją ocenę rozprawy, pragnę na wstępie podkreślić jej staranną redakcję i wysoki poziom językowy. Pragnę zwrócić uwagę na realizację w pracy zarówno tych aspektów rozwiązywanych problemów, które leżą w obszarze informatyki teoretycznej i statystycznej analizy danych, jak i aspektów praktycznych, obejmujących implementację i testowanie algorytmów w warunkach rzeczywistych. Do najważniejszych osiągnięć badawczych przedstawionych w rozprawie zaliczyłabym:
— Uwzględnienie w podejściu z rozdziału 5 dodatkowej informacji w postaci sieci Petriego z regułami towarzyszącymi tranzycjom. Doktorant najpierw ogranicza sieć procesów metabolicznych do fragmentu, który może wspomóc wyszukiwanie wzorców w danych transkryptomicznych, potem rozwiązuje dla niego problem optymalizacyjny sformułowany w postaci wyrażeń programowania liniowego całkowitoliczbowego mieszanego, a następnie uzyskany wektor aktywności stanów stosuje do wyodrębnienia charakterystycznych motywów w danych.
— Włączenie do opracowanej przez Doktoranta nowej wersji algorytmu RPCA nowego komponentu w postaci macierzy szumu, która zwiększyła znacznie skuteczność metody dekompozycji stosowanej do danych biologicznych.
— Przeprowadzenie badań na rzeczywistych danych biomedycznych i wykazanie, że zaproponowane algorytmy i schematy przetwarzania danych lepiej służą wydobywaniu istotnej informacji.

Jako pewien mankament odbieram brak w rozprawie porównania z wcześniejszymi podejściami metody zaproponowanej w rozdziale 2. Co prawda rozwiązywany problem jest nowy i nie istnieje inna metoda działająca przy tych samych założeniach, co usprawiedliwia brak takiego porównania, moim zdaniem można jednak było choćby fragmentarycznie odnieść się do wcześniejszych rozwiązań. Przykładowo, można było uruchomić algorytm dostosowany do danych pochodzących z heterogenicznego zbioru komórek i wykazać, że nie wykrył on wzorców obecnych w danych, podczas gdy nowa metoda to robi. W rozdziale 3 jednym z argumentów za nową metodą tRPCA był jej znacznie krótszy czas działania niż dla RPCA, czasów obliczeń jednak nie przytoczono. W moim odczuciu rozdział 5 można było pominąć, gdyż nie zawiera treści istotnych z punktu widzenia rozprawy z obszaru

informatyki. Z drugiej strony, rozumiem potrzebę zapoznania czytelnika z szerokim zakresem badań, w których Doktorant brał udział, oraz z jego biegłością w stosowaniu podejść i narzędzi matematycznych i informatycznych. Opis laboratoryjnego eksperymentu biologicznego w podrozdziale 2.5 uważam za zbyt szczegółowy. Z drobnych uwag natury edycyjnej: rozmiar czcionki na niektórych rysunkach jest zbyt mały (zwłaszcza na rys. 2.3, ale także 1.1, 5.1); w wyrażeniu $F(V, W, H)$ na str. 28 należałoby zastąpić V przez A; przykład na rys. 4.2 obejmuje sześć reakcji, nie pięć.

Powyższe uwagi nie podważają wartości rozprawy i mojej pozytywnej oceny pracy naukowej Doktoranta.

Ważnym aspektem badań Doktoranta było wyprowadzenie nowych, ważnych biologicznie wniosków poprzez zastosowanie zaproponowanych podejść algorytmicznych. Nie byłoby to możliwe bez dostępu do wartościowych i aktualnych danych pochodzących z eksperymentów biologicznych przeprowadzanych w instytutach medycznych. To z kolei nie byłoby możliwe bez aktywnej współpracy Doktoranta z badaczami z tych instytutów, w kraju i za granicą. Akceptacja przez stronę biomedyczną uzyskanych w ten sposób wyników, przedstawionych w rozprawie, potwierdzona publikacjami w liczących się czasopismach z listy JCR i cytowaniami, dodatkowo przemawia za wartością nowych propozycji.

3. Podsumowanie

Przedłożoną rozprawę uważam za napisaną w sposób przejrzysty, poprawny logicznie i metodologicznie. Pan mgr inż. Krzysztof Gogolewski wykazał się wiedzą i umiejętnościami w rozwiązywaniu problemów bioinformatycznych z użyciem podejść z zakresu matematyki i informatyki. Wykonał także olbrzymią pracę wejścia w celu dogłębnego zrozumienia biologicznego podłoża badań. Możliwość bezpośredniego zastosowania osiągniętych wyników badań przez stronę biologiczną uważam za ważny argument na korzyść tej pracy. Z czysto informatycznego punktu widzenia można zaobserwować biegłość Doktoranta w stosowaniu i projektowaniu algorytmów analizy danych, a także w ich implementowaniu i testowaniu na trudnych, wielowymiarowych i bardzo obszernych zestawach danych.

Doktorant potwierdził wagę swoich badań publikacjami przyjętymi m.in. w czasopismach z listy JCR. Posiada trzy takie publikacje za w sumie 80 punktów ministerialnych, które wg bazy *Web of Science* cytowane były 12 razy, a 21 razy wg *Google Scholar*.

Na podstawie wyrażonych powyżej opinii stwierdzam, że rozprawa pt. „Matrix methods in transcriptomic and metabolomic data analysis" autorstwa mgr. inż. Krzysztofa Gogolewskiego spełnia warunki stawiane rozprawom doktorskim przez obowiązującą ustawę o stopniach naukowych i tytule naukowym. Wnoszę o dopuszczenie tej rozprawy do publicznej obrony.

Marta Kasprzak