Prof. dr hab. inż. Andrzej Polański

Katedra Grafiki, Wizji Komputerowej i Systemów Cyfrowych

Politechniki Śląskiej

Gliwice 5.12.2023

Recenzja rozprawy doktorskiej

Autor: Mgr Agnieszka Geras

Tytuł: Modeling count data in Molecular Biology and Sociophysics: Selected Applications

Promotor: Prof. dr hab. Ewa Szczurek

Ogólna charakterystyka rozprawy

Przedłożona do recenzji rozprawa doktorska jest napisana w języku angielskim. Liczy 121 stron tekstu, składa się z siedmiu rozdziałów. Obszerna bibliografia liczy 145 pozycji. Tematyka pracy leży w obszarze inżynierii i analizy danych eksperymentalnych, pomiarowych lub obserwacyjnych. Jest to bardzo aktualny i silnie się rozwijający kierunek badań naukowych. Dlatego, mimo, że wybrane zastosowania proponowanych nowym metod modelowania są trochę rozproszone, to wybór tematyki pracy należy uznać za bardzo trafny i interesujący.

Praca ma bezpośrednie odniesienia do kilku opublikowanych artykułów oraz do kilku tekstów dostępnych w Internecie na platformie arxiv, których współautorką jest Doktorantka.

Pierwszy rozdział stanowi wstęp, a ostatni podsumowanie. Spośród pięciu rozdziałów analitycznych (2-6), trzy (trzeci, czwarty oraz szósty) stanowią bezpośrednie odniesienia do publikacji Doktorantki, a dwa (drugi i piąty) mają charakter stricte metodologiczny, są poświęcone definicjom oraz krótkim przedstawieniom/wyprowadzeniom modeli matematycznych oraz obliczeniowych wykorzystywanych w publikacjach Doktorantki. Należy docenić taką konstrukcję pracy. Rozdziały drugi i piąty orientują pracę w stronę dyscypliny naukowej Informatyka Techniczna i Telekomunikacja, z której Kandydatka się doktoryzuje. Ponadto wprowadzają metodologiczny punkt widzenia dla wyników uzyskanych w publikacjach Doktorantki, które mają charakter interdyscyplinarny.

Rozdział pierwszy stanowi wstęp do przedłożonej do recenzji rozprawy doktorskiej. Omawia się w nim obszary zastosowań proponowanych metod/algorytmów oraz przedstawia się problemy, których rozwiązywaniu poświęcone są rozwijane metody. W rozdziale tym, a także w pracy uwagę skupia się na dwóch głównych obszarach zastosowań.

Pierwszy obszar zastosowań to biologia molekularna, lub bardziej precyzyjnie transkryptomika oraz zastosowania bioinformatyki transkryptomice. Omawia się techniki sekwencjonowania transkryptomu pojedynczych komórek (single cell RNA sequencing sc-RNA-seq) oraz tkanek (bulk RNA sequencing). Formułuje się problem anotacji typów komórek na podstawie odczytów sc-RNA-seq. Przedstawia odniesienia do kilku, niedawno opublikowanych, cieszących się już dużym zainteresowaniem środowiska naukowego artykułów poświęconych temu zagadnieniu. Przedstawia się technikę transkryptomiki przestrzennej z odniesieniami do publikacji na ten temat. W następnej kolejności przestawiany jest problem dekompozycji typów komórek w danych tkankowych sekwencjonowania RNA oraz w danych transkryptomiki przestrzennej.

Drugi obszar zastosowań to sieci to analizy danych dotyczących cytowań lub negatywnych cytowań artykułów naukowych oraz danych dotyczących ocen odpowiedzi (postów) w serwisie Stack Exchange Q&A. Omawia się strukturę danych pochodzących z tych dwóch źródeł. Podkreśla się analogie pomiędzy nimi. Odpowiedzi w bazie Stack Exchange Q&A są (mogą być) oceniane przez dodanie do nich pozytywnych lub negatywnych ocen. Natomiast artykuły naukowe w klasycznej bibliometrii mogą być scharakteryzowane przez liczbę ich cytowań. Jednak możliwe jest, przez zastosowanie dość zaawansowanych metod analizy tekstów publikacji uzyskanie drugiego wymiaru charakteryzacji, negatywnych cytowań. Doktorantka odnosi się do odpowiedniej literatury naukowej. Cytowania oraz negatywne cytowania są analogiczne do pozytywnych i negatywnych ocen postów w bazie Stack Exchange Q&A.

Na końcu rozdziału pierwszego Doktorantka krótko omawia zawartość całej pracy.

W rozdziale drugim Doktorantka przedstawia aparat matematyczny potrzebny do projektów dopasowania modeli sieci Bayesowskich do danych transkryptomicznych typu scRNA ze strukturą przestrzenną oraz do profili sekwencjonowania RNA omawianych dalej w rozdziałach trzecim i czwartym. Definiuje pojęcie warunkowej niezależności oraz, bazując na tym opisuje model matematyczny sieci Bayesowskiej. Następnie opisuje procedurę próbkowania z modelu sieci Bayesowskiej. W kolejnym podpunkcie definiuje pojęcie „koca Markowa" („Markov blanket"). W tym miejscu występuje nieścisłość / błąd terminologiczny. W odniesieniu do zdefiniowanego pojęcia powinno się raczej stosowań nazwę/termin „otoczenie Markowa" („Markow boundary"). Taką samą krytykę można sformułować do tekstu publikacji [57], gdzie także definiuje się to pojęcie. W kolejnych podpunktach opisuje procedurę próbkowania bazującą na łańcuchach Markowa i algorytm Metropolisa – Hastingsa. Dalej opisywana jest procedura próbkowania Gibbsa, a następnie, w podpunkcie 2.3.4 oryginalna koncepcja połączenia procedur Metropolisa – Hastingsa i Gibbsa

dostosowana do próbkowania przestrzeni definiowanych przez parametry sieci Bayesowskich.

W rozdziale trzecim opisana jest teoria i narzędzie obliczeniowe „Celloscope" służące do dekompozycji przestrzennych danych transkryptomicznych. Narzędzie to bazuje na teorii i metodologii opisanej w rozdziale drugim. Wykorzystanie skonstruowanego narzędzia obliczeniowego pozwala na wyodrębnienie, na podstawie obrazu barwionego HE oraz przestrzennej transkryptomiki, kilkunastu typów komórek. W przykładowym zastosowaniu są to typy komórek występujące w mózgu myszy. Dla każdego pola w przestrzennych danych transkryptomicznych dopasowuje się model sieci Bayesowskiej, przedstawiony na rysunku 3.2. W sieci tej występuje obserwowana zmienna, Cgs opisująca ekspresję genu g w polu s. Dla zmiennej tej zakłada się ujemny rozkład dwumianowy, jego parametry zależą od kilku zmiennych ukrytych. Dodatkowo w modelu występuje zmienna Bgt – opisująca, zero-jedynkowo istnienie lub brak wpływu genu g na komórkę typu t. Istnienie tej zmiennej jest kluczową cechą (lub jedną z kluczowych cech) wprowadzonego modelu. Wydaje się, że dzięki zapisaniu w tabeli tej zmiennej odpowiednich, wynikających z wiedzy biologicznej (wyników eksperymentów pomiaru ekspresji i doboru odpowiednich poziomów odcięcia), danych, cały algorytm zyskuje stabilność i wiarygodność. Przestrzenie parametrów rozkładów ukrytych zmiennych są przeszukiwane z wykorzystaniem algorytmu MCMC złożonego z próbkowaniem Gibbsa, opisanego także w poprzednim rozdziale. Algorytm przeszukiwania prowadzi do uzyskiwania wysokich wartości wiarygodności sieci. Na bazie uzyskanych wartości parametrów dokonuje się podziału komórek na typy.

W następnej kolejności w rozdziale trzecim przedstawia się wyniki testowania / weryfikowania narzędzia „Celloscope". Najpierw weryfikacja obejmuje dane symulowane, dla których wykazuje się bardzo dobrą zgodność typowania z rzeczywistym stanem danych. Następnie przedstawia się wyniki wykorzystania narzędzia „Celloscope" dla danych przestrzennych ekspresji scRNA mózgu myszy. Wykazuje się, że udaje się odtworzyć znane struktury mózgu. Dokonuje się porównania narzędzia „Celloscope" z innymi podejściami i wykazuje się jego zalety. jako wynik wielu przeprowadzonych eksperymentów obliczeniowych przedstawia się rekomendacje dla doboru hiperparamterów sieci. Przedstawia się także wyniki badań wrażliwości algorytmu maksymalizacji wiarygodności sieci na dobór struktur macierzy Bgt. Konkretnie, zmieniane są wartości odcięcia i dla różnych macierzy prowadzi się dekompozycje. Porównanie wyników tych obliczeń prowadzi do konkluzji, że algorytm jest dość odporny na wartości poziomów odcięcia.

W rozdziale czwartym opisuje się narzędzie „ST assign", które jest pewnym rozwinięciem narzędzia „Celloscope" z poprzedniego rozdziału. Podobnie jak poprzednio, teoria i metodologia bazują na wynikach opisanych w rozdziale drugim. Narzędzie „ST assign" służy do jednoczesnej anotacji komórek na podstawie wyników sekwencjonowania sc-RNA-seq oraz dekompozycji składu mieszanin komórek z pomiarem ekspresji. Bazą do obliczeń jest model sieci Bayesowskiej przedstawiony na rysunku 4.1. W tym przypadku w

sieci Bayesowskiej występuję dwie obserwowane zmienne, pierwsza to podobnie jak poprzednio Cgs opisująca ekspresję genu g w polu s oraz dodatkowo druga Cgc opisująca ekspresję genu g w komórce c. Dla obu tych zmiennych tej zakłada się, że ich rozkłady są dane modelem ujemnego rozkładu dwumianowego. Parametry tych rozkładów zależą od zmiennych ukrytych. Dodatkowo, podobnie jak w poprzednim modelu, w sieci Bayesowskiej występuje zmienna Bgt – opisująca, zero-jedynkowo istnienie lub brak wpływu genu g na komórkę typu t. Bardziej złożona konstrukcja sieci pozwala na integrowanie danych dwóch typów, przestrzennych ekspresji RNA oraz ekspresji sc-RNA-seq pojedynczych komórek. Scenariusz eksperymentów obliczeniowych jest analogiczny jak w poprzednim rozdziale. Najpierw wykazuje się skuteczność opracowanego algorytmu dla danych symulowanych. Następnie przedstawia się wyniki analizy integracyjnej dwóch zbiorów danych, danych przestrzennych ekspresji RNA mózgu myszy (wykorzystywanych już w poprzednim rozdziale) oraz danych sc-RNA-seq komórek mózgu myszy. Uzyskuje się interesujące wyniki. Wyniki mają trochę charakter pilotażowy. Przedstawia się pewne trudności obliczeniowe wynikające ze złożoności danych oraz algorytmu. Nie porównuje się zaproponowanego algorytmu z literaturą (prawdopodobnie referencyjne algorytmy nie istnieją).

W rozdziale piątym przedstawia się rozkłady prawdopodobieństwa typu potęgowego, które są potem wykorzystywane, w rozdziale szóstym do modelowania danych dotyczących cytowań publikacji oraz pozytywnych i negatywnych ocen wpisów w serwisie Stack Exchange Q&A. Przedstawia się rozkłady prawdopodobieństwa liczby sąsiadów w bezskalowych grafach, rozkłady potęgowe, rozkład Zipfa oraz rozkład Tsalissa – Pareto.

W rozdziale szóstym przedstawia się analizy danych pochodzących z stron Stack Exchange Q&A oraz OpenCitations. Przedstawia się porównania obserwowanych liczb cytowań z rozkładem Tsalissa – Pareto. Wykazuje się także zgodności obserwowanej intensywności dołączania/dopisywania z prawami potęgowymi. Wreszcie analizuje się indeks Hirscha, zarówno dla bazy cytowań jak i dla bazy Stack Exchange. Analizuje się problem oceny/modelowania tego indeksu oraz bada się jaki wpływ mają negatywne oceny lub negatywne cytowania na jego wartości.

Ocena rozprawy

Moja ocena recenzowanej pracy doktorskiej jest bardzo pozytywna. Recenzowana praca jest bardzo ściśle związana z oryginalnymi publikacjami,

Geras, A., Darvish Shafighi, S., Domżał, K., Filipiuk, I., Rączkowska, A., Szymczak, P., ... & Szczurek, E. (2023). Celloscope: a probabilistic model for marker-gene-driven cell type deconvolution in spatial transcriptomics data. Genome Biology, 24(1), 120, w której przestawione są wyniki omawiane w rozdziale 3,

Geras, A., & Szczurek, E. (2023). ST-Assign: a probabilistic model for joint cell type identification in spatial transcriptomics and single-cell RNA sequencing data. bioRxiv, 2023-05, w której przedstawione są wyniki omawiane w rozdziale 4,

Geras, A., Siudem, G., & Gagolewski, M. (2020). Should we introduce a dislike button for academic articles?. Journal of the Association for Information Science and Technology, 71(2), 221-229, w której zamieszczone są wyniki omawiane w rozdziale szóstym.

Czasopisma naukowe, w których ukazały się te prace, Genome Biology oraz Journal of the Association for Information Science and Technology cieszą się dużą międzynarodową renomą oraz posiadają wysoką punktację ministerialną (odpowiednio 200 i 140 pkt). Artykuły Doktorantki w tych dwóch czasopismach naukowych, mimo krótkiego czasu od wydania były już kilkakrotnie cytowane.

We wszystkich wymienionych powyżej publikacjach Doktorantka jest pierwszą autorką. Oprócz trzech wymienionych powyżej prac Doktorantka jest współautorką kilku innych wysoko notowanych publikacji o tematyce zbieżnej z jej pracą doktorską.

Jak już wspomniano walorem pracy jest jej logiczna konstrukcja, poprzedzenie wyników eksperymentów i obserwacji opisami modeli matematycznych i obliczeniowych jest bardzo dobrym krokiem. Doktorantka dokumentuje opanowanie warsztatu naukowego obejmującego modelowanie matematyczne, prowadzenie projektów obliczeniowych.

Należy także podkreślić ładny i jasny styl pracy. W tekście jest wprowadzony jasny i precyzyjny formalizm matematyczny, co także zasługuje na uznanie. Konstrukcja algorytmu próbkowania łączącego iteracje MCMC oraz próbkowanie Gibbsa jest oryginalnym pomysłem pozwalającym na stochastyczną maksymalizację wiarygodności analizowanych sieci Bayesowskich.

Uwagi dyskusyjne i krytyczne

Mimo, że pracę bardzo dobrze się czyta, to jednak rozrzut tematyczny pomiędzy rozdziałami z zakresu bioifnormatyki oraz rozdziałami z zakresu analizy danych dotyczących cytowań artykułów naukowych oraz danych dotyczących postów w serwisie Stack Exchange Q&A trochę rzutuje na niejednorodność stosowanych metod i modeli.

Praca zyskałaby gdyby dołączyć do niej wykaz rysunków, tabel, listę skrótów.

Interesujące byłoby także zamieszczenie jakichś dodatkowych szczegółów implementacyjnych opracowanych algorytmów obliczeniowych, czasów obliczeń, kryteriów zatrzymania. Należy jednak podkreślić, że pseudokody stanowią już dobrą dokumentację. Oprogramowanie jest także dostępne na platformie Github.

Nasuwa się pytanie w jaki sposób, czy też na bazie jakich heurystyk były konstruowane sieci Bayesowskie opisane w rozdziałach trzecim i czwartym. Czy stosowano/wykonywano jakieś kroki prób i błędów? Jakimi przesłankami kierowano się wybierając postacie parametryczne rozkładów prawdopodobieństw w tych sieciach?

Czy możliwe jest, że opracowany i uruchomiony przez Doktorantkę algorytm stochastycznej maksymalizacji wiarygodności sieci Bayesowskiej doprowadzi do wykrycia lokalnego/fałszywego maksimum?

Konkluzja

Uwagi dyskusyjne i krytyczne nie podważają oryginalnych osiągnięć uzyskanych w pracy.

Praca stanowi podsumowanie oryginalnych publikacji naukowych, w których formułowano i weryfikowano oryginalne hipotezy badawcze. Osiągnięcia i oryginalne elementy rozprawy są na pewno wystarczające do jej ogólnej pozytywnej oceny. Stwierdzam, że rozprawa spełnia warunki stawiane pracom doktorskim i wnioskuję o jej dopuszczenie do publicznej obrony.