Wpł. 19.06.2023
RDITT – M. Skowy

Dr hab. inż. Grzegorz Dudek, prof. PCz
Katedra Automatyki, Elektrotechniki i Optoelektroniki
Wydział Elektryczny
Politechnika Częstochowska
Al. Armii Krajowej 17
42-200 Częstochowa

Częstochowa, dn. 16 czerwca 2023 r.

RECENZJA

rozprawy doktorskiej mgr inż. Aleksandry Suwalskiej

pt. *Developing a system of automatic identification of cellular subpopulations in data from single-cell mass cytometry with the use of algorithms for grouping of high dimensional data*

Formalną podstawą opracowania recenzji jest pismo Przewodniczącego Rady Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Śląskiej, prof. dr hab. inż. Andrzeja Polańskiego. Oceny rozprawy doktorskiej dokonano według kryteriów określonych w ustawie z 20 lipca 2018 r. *Prawo o szkolnictwie wyższym i nauce*. Promotorem rozprawy doktorskiej jest prof. dr hab. inż. Joanna Polańska.

Charakterystyka rozprawy

Rozprawa napisana jest w języku angielskim, liczy 109 stron. Składa się z sześciu rozdziałów (w tym dodatku), bibliografii zawierającej 59 pozycji, spisów tabel i rysunków oraz objaśnienia skrótów. Praca poprzedzona jest streszczeniami w języku polskim i angielskim.

Tezy pracy są następujące:

1. *The unsupervised machine learning methods allow data-driven identification of cell subpopulations leading to a better description of the cell-type heterogeneity.*
2. *Transformation of data feature space into an expanded representation using Gaussian Mixture Modelling allows for better separation and definition of the identified cell populations.*
3. *A two-step clustering approach with local feature space optimization enables the identification of rare cell populations.*

W rozdziale pierwszym, *Introduction to the mass cytometry data analysis*, przedstawiono motywację, cele i tezy pracy. Omówiono zagadnienia związane ściśle z tematyką rozprawy: zagrożenia i wyzwania związane z rozwojem gruźlicy, technologię cytometrii masowej, przygotowanie próbek, format zapisu danych cytometrycznych, schemat analizy danych, przetwarzanie danych, redukcję wymiarowości, próbkowanie i grupowanie danych. Omówiono też zbiory danych wykorzystywane w badaniach eksperymentalnych oraz metody wizualizacji.

Rozdział drugi, *Pre-gating of mass cytometry data*, omawia problem wstępnego bramkowania danych z cytometrii masowej. Scharakteryzowano zbiory danych oraz algorytm klasteryzacji oparty na mieszaninach gaussowskich. Przedstawiono i przedyskutowano wyniki eksperymentów.

W rozdziale trzecim, *Batch effect correction analysis*, omówiono problem korekty zakłóceń danych cytometrycznych. Przedstawiono procedurę badań z wykorzystaniem standardowych metod korekty zakłóceń, algorytmów grupowania i wizualizacji. Omówiono wyniki eksperymentów.

Rozdział czwarty, *Identification of cell subpopulations*, to najważniejszy i największy objętościowo rozdział rozprawy, w którym przedstawiono proponowaną metodę identyfikacji subpopulacji komórek na podstawie danych ze spektrometrii masowej. Scharakteryzowano istniejące metody grupowania stosowane do identyfikacji subpopulacji komórek. Zaprezentowano oryginalne podejście z rozszerzoną przestrzenią cech, selekcją cech i podwójnym grupowaniem danych. Opisano miary jakości grupowania. Przedstawiono i omówiono wyniki identyfikacji subpopulacji komórek uzyskane dla różnych wariantów problemu badawczego.

Badania podsumowano w rozdziale piątym.

Opinia na temat rozprawy - uwagi krytyczne i polemiczne

Cytometria masowa pozwala na badania pojedynczych komórek, przyczyniając się do dogłębnego zrozumienia procesów komórkowych oraz identyfikacji nowych subpopulacji komórek i ich roli w organizmie. Dzięki analizie danych cytometrycznych, które obejmują poziomy ekspresji dla każdej komórki i markera, można identyfikować podtypy komórek, różnice w aktywności genów oraz interakcje międzykomórkowe. Wiedza zdobyta dzięki cytometrii masowej niesie ze sobą ogromne możliwości zastosowań praktycznych. Otwiera nowe perspektywy w dziedzinie rozwoju innowacyjnych terapii, leków oraz strategii zapobiegania i leczenia różnorodnych schorzeń.

Cytometria masowa jako zaawansowana technologia analizy pojedynczych komórek generuje wielowymiarowe dane składające się z milionów komórek i kilkudziesięciu cech. Identyfikacja podtypów komórek w kontekście tak dużej ilości danych stanowi wyzwanie. Stawiając czoła temu wyzwaniu, Autorka opracowała schemat przetwarzania i analizy danych pozyskanych za pomocą cytometrii masowej, obejmujący procedury oczyszczania i selekcji danych, redukcji błędów związanych z metodami pozyskiwania danych, wzbogacania i przetwarzania przestrzeni cech oraz grupowania i wizualizacji wielowymiarowych danych. Stosując ten schemat i wykorzystując nowoczesne metody analizy danych, zidentyfikowała subpopulacje komórek w dwóch zbiorach danych. Jeden z tych zbiorów obejmuje komórki pacjentów chorych na gruźlicę. Wyniki pracy są bardzo cenne. Po pierwsze, przyczyniają się do poznania składu i funkcji komórek pacjentów z gruźlicą, co może pomóc w opracowaniu nowych metod leczenia i profilaktyki. Po drugie, opracowany schemat pozwala zautomatyzować pracę z wielowymiarowymi danymi pozyskanymi za pomocą cytometrii masowej. Tematykę pracy uznaję za bardzo ważną, społecznie pożyteczną i aktualną w aspekcie jej walorów poznawczych i utylitarnych.

Tytuł rozprawy jest odpowiednio zwarty i komunikatywny. W pełni oddaje najistotniejsze elementy treściowe rozprawy. Motywacja do podjęcia badań przedstawiona przez Autorkę w rozdziale pierwszym jest przekonująca. Problem badawczy jest jasno sformułowany, a cel badań jest czytelny: opracowanie procedury grupowania dużych zbiorów danych cytometrycznych wykazującej dużą czułość na małe subpopulacje komórek, uwzględniającej heterogeniczność danych biologicznych i automatyzującej procesy oczyszczania danych i redukcji wariancji technicznej. Tezy pracy są poprawne, oryginalne i jednoznaczne, a przy tym spójne z założonym celem. Można by je jednak uściślić, dodając informację, że badania odnoszą się do danych pozyskanych za pomocą cytometrii masowej.

We podrozdziale 1.4 (*Background*) Autorka przygotowuje czytelnika do lektury kolejnych rozdziałów, definiując kluczowe zagadnienia. Opisuje zagrożenia i wyzwania związane z rozwojem gruźlicy, uzasadniając swoje badania eksperymentalne z wykorzystaniem danych pacjentów chorych na tę chorobę. Skrupulatnie omawia technologię cytometrii masowej – przygotowanie próbek i schemat działania cytometru masowego. W kolejnym podrozdziale przedstawia format zapisu wyników badania, obrazując go rysunkiem 1.2. Ten rysunek wymaga jednak komentarza: co oznaczają „pushes", jakie znaczenie ma „signal treshold" i „event not counted"? Podrozdział 1.4.6 zawiera standardowy schemat analizy danych cytometrycznych. Kolejne kroki tej analizy scharakteryzowano w następnych podrozdziałach. Ta charakterystyka wymaga jednak doprecyzowania. Nie jest jasne jak przeprowadza się normalizację (*bead normalization*), co oznacza zdanie „raw integers or non-zero integers randomized between ranges [x-1, x]", zastosowanie transformacji typu *logicle* i *ArcSinh* wymaga dokładniejszego komentarza, podobnie jak procedury kompensacji i *debarcoding*. W kolejnych podrozdziałach opisano problem redukcji wymiarowości, próbkowania i grupowania. Autorka porównuje standardowe algorytmy redukcji wymiarowości (PCA, t-SNE i UMAP), dyskutuje próbkowanie danych cytometrycznych i charakteryzuje kilka standardowych algorytmów grupowania danych: k-średnich, metodę oparta na mieszaninach gaussowskich i algorytm aglomeracyjny.

W podrozdziale 1.5 opisano dane używane w badaniach eksperymentalnych, a w podrozdziale 1.6 przedstawiono metodę redukcji wymiarowości UMAP i wizualizacji danych mISO. Charakterystyka tych metod jest bardzo oszczędna. Jako podstawowe narzędzie wizualnej oceny wyników badań, metoda UPMAP powinna być szczegółowo opisana. Przyjęte wartości parametrów tej metody (n_neighbors=30, min_dist=0.2) powinny zostać uzasadnione. Czy zbadano wpływ tych parametrów na wyniki?

W rozdziale drugim przedstawiono automatyczną procedurę wykrywania i usuwania „wadliwych" obserwacji ze zbioru danych, np. martwych lub zdublowanych komórek. Ta nowatorska procedura wykorzystuje mieszaninę dwuwymiarowych rozkładów gaussowskich, które stosuje się do wykrywania skupisk w danych skomponowanych z wybranych dwóch markerów. Każda para markerów wykorzystywana jest do detekcji innego typu „nieprawidłowości". Dla poprawy czytelności sugerowałbym zwizualizowanie algorytmu klasteryzacji opisanego w podrozdziale 2.2.1. Wyjaśnienia wymaga przyjęty 15-procentowy próg akceptacji – czy ta wartość będzie odpowiednia dla innych zbiorów danych? Wyniki badań zostały szczegółowo przedyskutowane. Podkreślono zalety metody – automatyzację, powtarzalność wyników, niezależność od wielkości próby, brak próbkowania (które może usuwać rzadkie subpopulacje komórek) oraz możliwość rozszerzenia o dodatkowe kryteria

filtracji. Zaproponowane podejście opiera się na grupowaniu z wykorzystaniem mieszanin gaussowskich. W pracy zabrakło szczegółowego opisu tego algorytmu.

Rozdział trzeci dotyczy korekcji wariancji technicznej danych z cytometrii masowej. Po dyskusji różnych metod redukcji wariancji, Autorka wybiera te najbardziej odpowiednie do specyfiki jej danych, iMUBAC i cyCombine. Wykorzystując te metody, dokonuje korekty wariancji i identyfikuje subpopulacje komórek za pomocą grafowej metody klasteryzacji PARC i algorytmu Leiden. Porównuje wyniki z tymi otrzymanymi dla danych bez korekty. Dodatkowo wyniki obu metod, iMUBAC i cyCombine, grupuje za pomocą algorytmu aglomeracyjnego, aby potwierdzić utworzenie równoważnych grup w obu podejściach. Do wizualizacji wyników przed i po korekcie wykorzystano metodę UMAP i mISO. Dzięki wizualizacji można zaobserwować zmiany położenia poszczególnych komórek przed wykonaniem i po wykonaniu korekty. Ciekawym rozwiązaniem jest nauczenie sieci neuronowej transformacji UMAP. Pozwala to zastosować tę transformację do nowych komórek. Po gruntownej dyskusji wyników, Autorka rekomenduje metodę cyCombine jako bardziej efektywną od iMUBAC dla rozważanego zbioru danych. Badania zrelacjonowane w rozdziale drugim i trzecim dowodzą tego, że Autorka dobrze orientuje się w metodach analizy i przetwarzania danych cytometrycznych i swobodnie operuje tymi metodami, aplikując je do rozwiązania postawionego problemu. Szkoda, że algorytmy te nie zostały szczegółowo opisane, co pozwoliłoby czytelnikowi zrozumieć ich naturę.

W rozdziale czwartym opisano proponowaną metodę identyfikacji subpopulacji komórek. Na początku tego rozdziału Autorka charakteryzuje znane algorytmy identyfikacji subpopulacji komórek oparte na grupowaniu danych. Dyskutuje ich ograniczenia związane głównie ze słabą skalowalnością, potrzebą próbkowania danych skutkującą pomijaniem małych subpopulacji, stochastyczną naturą, potrzebą podania liczby klastrów a priori i zaangażowania ekspertów do wstępnej analizy danych. Analizując zbiór danych, Autorka zauważa wysoka heterogeniczność, która utrudnia poprawną identyfikacją subpopulacji. Do wykrywania subpopulacji stosuje metodę mieszanin gaussowskich. Aby poprawić zdolność algorytmu do identyfikacji niewielkich subpopulacji proponuje rozszerzyć przestrzeń cech. Dodatkowe cechy powstają na bazie prawdopodobieństw warunkowych wyznaczonych ze składowych mieszaniny gaussowskiej. Prawdopodobieństwa te zamienione zostają na stopnie przynależności, aby uniknąć problemów wprowadzanych przez wielomodalność rozkładów tych prawdopodobieństw. Rozważa się też wersję binarną wygenerowanych nowych cech. Autorka zaproponowała metodę selekcji cech opartą na ich statystykach i aproksymacji histogramów tych statystyk mieszaniną rozkładów gaussowskich. Zauważając, że rozszerzenie i optymalizacja przestrzeni cech może nie wystarczyć do ujawnienia rzadkich subpopulacji komórek, Autorka proponuje dodatkową klasteryzację wyłonionych wcześniej grup za pomocą dwóch algorytmów: PARC i k-średnich. Do oceny jakości grupowania zaproponowano kilka miar. Nie podano jednak wzorów definiujących najważniejsze z tych miar: wskaźnik Calińskiego-Harabasza i wskaźnik Daviesa-Bouldin'a. Wyniki identyfikacji subpopulacji komórek dla dwóch zbiorów danych są bogato ilustrowane i komentowane przez Autorkę.

Dyskutując rezultaty badań w podrozdziale 4.5, Autorka trafnie zauważa, że heterogeniczność danych może znacząco wpłynąć na wyniki grupowania i zakłócić obraz subpopulacji komórek przez co ich identyfikacja dokonywana „ręcznie" przez ekspertów budzi duże wątpliwości. Zaproponowana przez Autorkę automatyczna procedura identyfikacji, dzięki wbudowanym mechanizmom generacji i selekcji

cech, pozwala zredukować wpływ heterogeniczności na rozpoznanie subpopulacji. Umożliwia przetwarzanie dużych zbiorów danych i detekcję niewielkich subpopulacji. Autorka potrafi spojrzeć krytycznie na zaproponowane przez siebie podejście i dostrzega jego ograniczenia związane głównie z dużą złożonością obliczeniową, trudnościami w doborze parametrów i nieoptymalnym podziałem na subpopulacje. Wskazuje również na możliwość dalszego rozwoju, szczególnie poprzez zastosowanie metod uczenia głębokiego. Praca została podsumowana w rozdziale piątym.

Sekwencja treści prezentowych w kolejnych rozdziałach pracy jest właściwa: od informacji wstępnych, wyjaśniających zagadnienia poruszane w pracy, poprzez opis danych wykorzystywanych w części eksperymentalnej, opis badań wstępnych mających na celu przygotowanie danych do dalszej analizy, po omówienie proponowanej metody identyfikacji subpopulacji komórek i analizę wyników. Praca napisana jest starannie, poprawnym językiem naukowym z właściwym słownictwem specjalistycznym i odpowiednią ścisłością sformułowań. Układ redakcyjny rozprawy nie budzi zastrzeżeń, z dwiema uwagami. Motywacja, cel i tezy pracy można było zamieścić w odrębnym rozdziale, a nie w rozdziale pt. *Introduction to the mass cytometry data analysis* razem z opisem kluczowych zagadnień, danych i algorytmów. Z podrozdziału 4.1 niepotrzebnie wyodrębniono podrozdział 4.1.1 (tylko jeden).

Niedostatkiem pracy jest brak szczegółowych opisów metod składowych proponowanego rozwiązania: metody grupowania z wykorzystaniem mieszanin gaussowskich, UMAP, PARC, algorytmu Leiden itp. Brakuje też definicji stosowanych miar i statystyk: czynnika Bayesa, statystyki GAP, wskaźnika Calińskiego-Harabasza i wskaźnika Daviesa-Bouldin'a.

Tezy rozprawy zostały udowodnione. Autorka wykazała, że zastosowane metody uczenia nienadzorowanego umożliwiają identyfikację subpopulacji komórek na podstawie danych z cytometrii masowej, co prowadzi do lepszego opisu heterogeniczności różnych typów komórek. Rozszerzona reprezentacja danych komórkowych otrzymana przy użyciu mieszanin gaussowskich pozwala na dokładniejszą identyfikację subpopulacji komórek niż metody standardowe. Dwuetapowe podejście do grupowania z lokalną optymalizacją przestrzeni cech umożliwia identyfikację niewielkich subpopulacji komórek. Nie mam wątpliwości, że Autorka rozwiązała postawiony problem badawczy i osiągnęła założone cele, używając właściwych metod. Oryginalność rozprawy polega na opracowaniu metodyki identyfikacji subpopulacji komórek na podstawie danych z cytometrii masowej. Rozprawa wykorzystuje aktualne osiągnięcia w dziedzinie biologii molekularnej i w dziedzinie analizy danych biologicznych opisane w literaturze światowej. Autorka wykazała umiejętność poprawnego i przekonywującego przedstawienia i interpretacji uzyskanych wyników. Na pochwałę zasługuje jej wnikliwość i drobiazgowość w analizie wyników badań i formułowaniu wniosków. Oceniam wysoko znaczenie uzyskanych wyników dla rozwoju dyscypliny informatyka techniczna i telekomunikacja, zwłaszcza w kontekście ich potencjalnych zastosowań w biologii.

Uwagi językowe, edytorskie i redakcyjne

- W streszczeniu w języku polskim występują błędy stylistyczne: „Wybuch pandemii COVID-19 doprowadził do regresji w postępach ostatnich lat w powstrzymywaniu roznoszenia się gruźlicy", „Naukowcy próbują zmniejszyć występowanie nowo zdiagnozowanych przypadków do

minimalnego poziomu poprzez pracę nad nowymi terapiami i lekami", „Publicznie dostępne zbiory danych posłużyły do zaproponowania schematu analizy, który rozwiązuje problemy występujące przy istniejących rozwiązaniach".

- Str. 22: we wzorze (2.2) występują symbole a_1 i a_2, a w tekście symbole α_1 i α_2.
- Str. 32: brak objaśnienia symbolu "k" we wzorze (3.1).
- Str. 32: Powinno być „Wilcoxon" zamiast „Wilcovon".
- Str. 59: brak objaśnienia symbolu "E[RI]" we wzorze (4.2).

Wniosek końcowy

Zakres tematyczny rozprawy doktorskiej mgr inż. Aleksandry Suwalskiej i osiągnięte w niej oryginalne wyniki lokują tę rozprawę w obszarze zastosowań współczesnej informatyki w dziedzinie nauk biologicznych. Uważam, że rozprawa stanowi oryginalne rozwiązanie problemu naukowego i wskazuje na wysoki poziom wiedzy Autorki w zakresie dyscypliny informatyka techniczna i telekomunikacja, a także na umiejętność samodzielnego prowadzenia przez nią badań naukowych.

Stwierdzam, że opiniowana rozprawa doktorska spełnia wymogi ustawy z 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce. Wnoszę o dopuszczenie mgr inż. Aleksandry Suwalskiej do publicznej *obrony pracy doktorskiej.*