Dr hab. inż. Paweł Piotr Łabaj
Małopolskie Centrum Biotechnologii
Uniwersytet Jagielloński
Gronostajowa 7a, 30-387 Kraków
pawel.labaj@uj.edu.pl

Kraków, 03.11.2022

Recenzja rozprawy doktorskiej

Tytuł rozprawy:	Clustering techniques of high-throughput big -omics data
Autor rozprawy:	mgr inż. Grzegorz Mrukwa
Promotor rozprawy:	Prof. dr hab. inż. Joanna Polańska
Dziedzina:	nauki techniczne
Dyscyplina:	Informatyka techniczna i telekomunikacja

Ostatnie lata w naukach biologicznych i medycznych cechują się szybkim rozwojem różnych biomedycznych technik pomiarowych jak i poszerzaniem obszarów ich wykorzystania. Wzrost ilości mierzonych, przechowywanych i przetwarzanych danych dotyczy nie tylko objętości, ale też rozdzielczości i złożoności. Jedną z nich jest obrazowanie z wykorzystaniem spektrometrii mas (Mass Spectrometry Imaging, MSI), które pozwala na kompleksowe pozyskiwanie danych dotyczących dystrybucji przestrzennej związków chemicznych na obszarze badanej tkanki. Technologia ta umożliwia między innymi szczegółowe badania nad klonalnością nowotworów, analizą ich pochodzenia, typów i podtypów. Wyzwaniem jest jednak właściwa ocena uzyskanych obrazów, zwłaszcza przy niewystarczającej ilości ekspertów.

W obliczu szybkiego rozwoju algorytmów z obszaru „sztucznej inteligencji" / „uczenia maszynowego" otwiera się okno na pełne wykorzystanie potencjału technologii takich jak MSI. W przedstawionej rozprawie doktorskiej Autor adresuję problem nienadzorowanej eksploracji danych pochodzących z eksperymentów MSI. A mianowicie zauważa, że istniejące podejścia do grupowania danych MSI są zwykle oparte o projekcję wysokowymiarowego zbioru obserwacji do przestrzeni niskowymiarowej, a następnie stosowany jest klasyczny algorytm klasteryzacji. Niestety stosowane rozwiązania jednokrokowe mają ograniczoną czułość spowodowaną przez pomijanie lokalnych niuansów dystrybucji danych przez metody projekcji. Autor proponuje jako rozwiązanie tego wyzwania zastosowanie podejścia hierarchicznego, gdzie zamiast jednokrotnej transformacji danych do przestrzeni niskowymiarowej, wielokrotnie wykonywana jest selekcja cech, ale ograniczona jedynie do aktualnie przetwarzanego obszaru. Takie podejście ma pozwolić na uniknięcie ograniczeń wynikających z globalnej projekcji danych. W oryginalnie zaproponowanym podejściu – DiviK – wykorzystywana jest metoda

filtracji cech z użyciem modelu mieszanin Gaussa i metodę k-średnich skalibrowanej na potrzeby danych MSI, natomiast w jego rozwinięciu – DiVAE – w opracowanym schemacie osadzony został algorytm oparty o uczenie głębokie zaproponowany wcześniej przez innego badacza. Jak zauważa Autor: *„Podejście to otwiera nowe możliwości również w przetwarzaniu innych rodzajów danych biologicznych, gdyż zastosowanie grupowania głębokiego zwykle wymaga jedynie dostosowania architektury sieci neuronowej do dystrybucji danych w nowej dziedzinie"*, co jest zgodne z prawdą i powoduje, że z uwagą będę śledził dalsze losy Autora jak i dalszy rozwój metod DiviK i DiVAE.

Układ rozprawy jest typowy dla tego typu opracowań. W kolejnych rozdziałach Autor:

- opisuje technologię MSI i wyjaśnia pochodzenie danych użytych w pracy jak również konsekwencje wybranej procedury pozyskiwania na jakość danych (Rozdział 2),
- przybliża obecny stan wiedzy w zakresie klasteryzacji *big-omics data* (Rodział 3),
- przedstawia autorską metodę *Divisive Intelligent K-Means* (DiviK) oraz wyniki porównania z innymi metodami (Rozdział 4),
- przedstawia kolejny etap rozwoju DiviK a mianowicie *Divisive clustering with Variational Autoencoders* (DiVAE) oraz wyniki porównania (Rozdział 5).

W ostatnim rozdziale następuje podsumowanie pracy wraz z wskazaniem potencjalnych obszarów zastosowania oraz zarysowaniem dalszych kierunków rozwoju.

Autor szeroko i bardzo szczegółowo omawia obecny stan wiedzy co świadczy o dużym oczytaniu i dobrym przygotowaniu do podjęcia zagadnień poruszanych w dalszych częściach pracy. Jednakże, w wielu wypadkach zbyt szczegółowe opisy prowadzą do zaciemnienia obrazu i czytelnik może stracić z oczu właściwy cel pracy. Odnosi się wrażenie, że celem nie jest przybliżenie metody DiviK i jej rozszerzenia DiVAE, a jedynie dokonanie bardzo szczegółowego przeglądu dostępnych rozwiązań. Oczywiście przybliżenie stanu wiedzy jest kluczowe aby wskazać wyzwania i przedstawić jak proponowane podejście sobie z nimi radzi, ale w proponowanej pracy ten element jest przeładowany. Znacznie lepiej sprawdziłoby się przedstawienie klas rozwiązań z bardziej szczegółowym opisem jedynie tych podejść, które ostatecznie będą użyte w porównaniu. Brakuje też jasnego zaznaczenia, które metody w porównaniu będą użyte.

Należy jednak nadmienić, że tezy rozprawy są sformułowane jasno i przystępnie oraz są w pełni poparte danymi zawartymi w poszczególnych rozdziałach rozprawy. Podsumowanie rozprawy jest syntetycznym wykazaniem, że założone w pracy cele zostały osiągnięte. Doktorant dla każdej z tez wskazuje, która część rozprawy się do niej odnosi. Głównym niedociągnięciem pracy jest wspomniane wyżej nieumiejętne przeprowadzenie czytelnika przez główne części rozprawy tak aby nie stracił z oczu jej celu. Jednocześnie Autor nadużywa sformułowania „uważny czytelnik", nie za bardzo rozumiem cel takiego zabiegu. To co mnie osobiście razi jest używanie liczby mnogiej „my" w rozprawie, która jednak dotyczy autorskiej metody. Oczywiście rozwijanej pod nadzorem, ale z głównym wkładem Autora. W tym kontekście zastanawiam się również

dlaczego nie zacytowano pracy o DiviK (https://doi.org/10.48550/arXiv.2009.10706), która wprawdzie nie ukazała się jeszcze w recenzowanym czasopiśmie, ale wśród piśmiennictwa znalazły się prace o podobnym statusie. Kończąc ten wątek, brakuje mi też krótkiego podsumowania dotychczasowych artykułów Autora w kontekście tematu pracy.

Rozprawa zawiera 27 szczegółowych rycin oraz 10 tabel, które co do zasady są jasne i czytelne. Jedyne zarzuty można mieć odnośnie ryciny 4.1 przedstawiającej ideę podejścia DiviK. W mojej ocenie nie oddaje ona jasno koncepcji. „Uważny czytelnik" może zadać pytanie czym właściwie różni się krok 1 od kroku 2. Ponadto w tabelach z wynikami 4.1, 4.2, i 4.3 kolejność wierszy wydaje się być zgodna z „overall quality" z Tabeli 4.3. Niestety nigdzie przy odniesieniach do tych tabel lub w ich opisach nie ma takiej informacji co utrudnia zrozumienie logiki prezentowanych wyników. Rozprawa zawiera też 10 równań, a piśmiennictwo obejmuje 140 dobrze dobranych i aktualnych pozycji.

Podsumowując, pomimo pewnych niedociągnięć, przedstawiona do oceny praca doktorska stanowi bardzo ciekawe i wartościowe rozwiązanie ważnego zagadnienia naukowego. Rozprawa jest ważnym przyczynkiem w zakresie wiedzy na temat klasteryzacji *big-omics data* a przedstawione podejście jak i kierunek rozwoju może mieć znaczny wpływ na dziedzinę. Praca ta w pełni odpowiada warunkom stawianym rozprawom doktorskim oraz wypełnia istotną lukę, w zakresie praktycznej wiedzy z zakresu technik informatycznych, warunkującej powodzenie analizy danych z eksperymentów biologicznych i medycznych. Jednak należy zauważyć, że możliwość zastosowania zaproponowanego rozwiązania wykracza daleko poza nauki biologiczne i medyczne.

Na podstawie powyższej oceny wnioskuję zatem do Wysokiej Rady Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Śląskiej o dopuszczenie Doktoranta do dalszych etapów przewodu doktorskiego. Nie mam wątpliwości, że doświadczenie zgromadzone przez Autora stawia cały zespół badawczy w doskonałej pozycji wśród międzynarodowych grup zajmujących się tą tematyką.

Łabaj Paweł

Dr hab. inż. Paweł Piotr Łabaj