Wydział Matematyki
i Nauk Informacyjnych

Warszawa, 20 grudnia 2022

Dr hab. inż. Przemysław Biecek, prof. uczelni
Wydział Matematyki i Nauk Informacyjnych
Politechnika Warszawska
Koszykowa 75/507
00-662 Warszawa

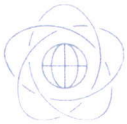
Recenzja rozprawy doktorskiej „Clustering techniques of high-throughput big-omics data" pana mgr. inż. Grzegorza Mrukwy.

Promotor rozprawy: prof. dr hab. inż. Joanna Polańska.

Promotor pomocniczy: dr inż. Michał Marczyk.

Niniejsza recenzja oparta jest o otrzymaną rozprawę doktorską oraz kody źródłowe załączone do rozprawy.

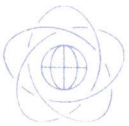
Główny cel rozprawy doktorskiej jest sformułowany w pierwszym rozdziale w następujący sposób: „provide a consistent and scalable framework for knowledge discovery in different kinds of big -omics data, with a specific focus on Mass Spectrometry Imaging". Tak postawiony cel już na początku



sugeruje narzędziowy charakter rozprawy. Zaproponowane w rozprawie rozwiązanie to zbiór metod wspierających grupowanie (klastrowanie) obserwacji oraz ich implementacja w bibliotece divik dla języka Python. Zbudowane rozwiązanie jest zastosowane do analizy dwóch zbiorów danych dotyczących obrazowania spektrometrii masowej, mianowicie "Oral Squamous Cell Carcinoma Dataset" (OSCC) i "Mouse Kidney 3D Dataset" (MK3D). Tak postawiony cel autor zrealizował. Złożona rozprawa to opis kompletnego rozwiązania wraz z prezentacją działania na dwóch zbiorach danych oraz omówienie otrzymanych wyników.

Struktura i zawartość pracy jest bezpośrednio związana z celem rozprawy. Rozdział drugi 'Material' składa się z trzech podrozdziałów. Pierwszy opisuje krótko wyzwania związane z analizą danych pochodzących z obrazowania spektrometrii masowej (ang. Mass Spectrometry Imaging, MSI). Tymi wyzwaniami są skalowalność oraz dobór właściwych procedur przetwarzania wstępnego, jak np. normalizacja czy usuwanie tła. Pozostałe dwa podrozdziały opisują zbiory danych OSCC i MK3D.

Rozdział trzeci 'Methodology of Big -omics Data Clustering' omawia metody wykorzystywane w analizie danych MSI. Ten rozdział składa się z pięciu podrozdziałów i podsumowania. Kolejne podrozdziały przedstawiają metody inżynierii cech dla danych MSI, bardzo krótkiego (8 linijek) podrozdziału dotyczącego kryteriów oceny algorytmów do klastrowania, podrozdziału przedstawiającego klasyczne algorytmy klastrowania w zastosowaniu do analizy danych MSI, podrozdziału przedstawiającego podejścia do kastrowania uwzględniające informacje przestrzenną obecną w danych MSI, oraz podejścia do klastrowania oparte o sieci głębokie. Konkluzja, moim zdaniem słuszna, dotyczy obserwacji, że w klasycznych metodach nie wykorzystuje się w pełni informacji o niskowymiarowej reprezentacji klastrowanych danych. W tym sensie ważność cech, na potrzeby inżynierii cech lub klastrowania, jest globalna i być może niezbyt adekwatna do faktycznie istotnych ukrytych charakterystyk kastrowanych obserwacji. Ten wniosek jest punktem wyjścia do pracy nad sprytniejszą reprezentacją danych, która zapewniałaby skalowalność, ale jednocześnie byłaby bardziej elastyczna niż podejścia klasyczne.



Rozdział czwarty 'Divisive Intelligent K-Means' opisuje rozwiazanie odpowiadające na wymienione wcześniej potrzeby i jest to główna kontrybucja recenzowanej pracy doktorskiej. Składa się z trzech głównych podrozdziałów poprzedzonych wstępem i zakończonych podsumowaniem. Drugi podrozdział opisuje dokładniej składowe rozwiązania czyli sposób filtrowania cech, sposób klastrowania i kryterium stopu w podziale klastrów. Trzeci podrozdział mierzy się z problemem skalowalności, a czwarty przedstawia wyniki zastosowania algorytmu DiviK do prawdziwych zbiorów danych OSCC i MK3D.

Rozdział piąty 'Divisive Clustering via Variational Autoencoders' wprowadza nowy pomysł na wykorzystanie autoenkoderów do tworzenia zredukowanej reprezentacji danych, która następnie może być użyta do klastrowania obserwacji klasycznymi metodami. Sam pomysł jest bardzo ciekawy i obiecujący, ale jeszcze wymaga pogłębionej analizy właściwości jakie ma ta nowa reprezentacja. Traktuję go więc jako bardzo obiecujące badania wstępne w tej tematyce, które z pewnością warto rozwijać z uwagi na korzystne właściwości autoenkoderów, takie jak możliwość zaaplikowania raz wytrenowanej reprezentacji do nowych danych. Z punktu widzenia metodologicznego ten rozdział ma największy potencjał na jakościowo nowe wyniki w odniesieniu do analiz danych MSI.

Rozprawę kończy podsumowanie i bibliografia. Ta ostatnia składa się ze 140 pozycji, głównie artykułów naukowych opublikowanych w ostatnich latach.

Warto podkreślić, że algorytm DiviK jest dostępny w postaci pakietu oprogramowania dla języka Python dostępnego na otwartej licencji Apache na stronie https://github.com/gmrukwa/divik oraz w serwisie PyPi. Sądząc po statystykach z serwisu GitHub prace programistyczne nad tym pakietem toczyły się głównie w latach 2018 i 2019 i zakończyły się preprintem opublikowanym w roku 2020 na serwisie arxiv (https://arxiv.org/abs/2009.10706). Preprint ten końcowo został opublikowany w listopadzie 2022 w czasopiśmie BMC Bioinformatics (https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-022-05093-z). Jeżeli chodzi o algorytm Di-

VAE to nie udało mi się znaleźć dostępnej publicznie implementacji tego algorytmu, preprintu ani publikacji w recenzowanym czasopiśmie.

Bazując na powyższym, pozostaje mi uznać, że głównym wynikiem doktoratu jest projekt, implementacja i przykłady użycia algorytmu DiviK. To rodzi jedną z wątpliwości, do których chciałbym by doktorant się ustosunkował na obronie, mianowicie związku rozprawy doktorskiej z artykułem opublikowanym w czasopiśmie BMC Bioinformatics. Pomimo iż jest bardzo duże przecięcie tych dwóch dzieł (np. rycina 1 z artykułu to rycina 4.1 z rozprawy, rycina 2 z artykułu to rycina 4.2 z rozprawy, rycina 4 z artykułu to rycina 4.4 z rozprawy itp), to nie ma w rozprawie żadnego odniesienia do tej publikacji. Publikacja w BMC Bioinformatics została opublikowana po złożeniu rozprawy, ale preprint był dostępny od roku 2020 i autor mógł i powinien się do niego odnieść we wstępie do rozprawy. Aby uniknąć nieporozumień co do określenia względnych relacji pomiędzy rozprawą a artykułem proponuję autorowi rozszerzyć wstęp do rozprawy wyraźnie zaznaczając, które wyniki zostały niezależnie opublikowane w postaci artykułu naukowego, a które nie są nigdzie indziej opublikowane wraz z informacją czy są podstawą do planowanych publikacji czy nie.

Główną zaletą i (w mojej ocenie) wartością rozprawy jest zaproponowanie rozwiązania metodologicznego i programistycznego do analizy danych MSI. Tym samym główny niedosyt wynika z braku informacji na temat siły oddziaływania tak opracowanego rozwiązania. Czy jest ono używane przez inne zespoły badawcze? Czy jest ono utrzymywane? Jakie są dalsze plany co do rozwoju tego rozwiązania. Rozdział 6.3 sugeruje, że autor większą nadzieję pokłada jednak w podejściach inspirowanych uczeniem głębokim, takimi jak choćby autoenkodery. Poza jednak krótką dyskusją w rozdziale 5.5, brakuje w rozprawie bardziej wyraźnej oceny potencjału użycia metody DiviK w przyszłości. W latach 2018-2020 autor rozprawy był współautorem wielu prac naukowych, też poświęconych tematyce MSI. Jestem więc przekonany, że odnosząc się też do wyników z tych prac mógłby napisać znacznie więcej o potencjale użycia algorytmu DiviK.

Politechnika
Warszawska

ul. Koszykowa 75
00-662 Warszawa

Głównym brakiem jaki odczuwam w przedstawionej rozprawie jest nieobecność pogłębionej dyskusji dotyczącej zagadnień bezpieczeństwa i stabilności dla zaproponowanych metod. Literatura z obszaru uczenia maszynowego przeżywa prawdziwy zalew przykładami modeli, które pomimo dobrych wyników na danych treningowych nie osiągają zadowalających wyników na danych testowych. Rośnie liczba głosów wymagających by coraz większa automatyzacja procesu analizy danych szła w parze z pogłębioną analizą odporności na zjawiska typu data-drift, robustness, bias, transparency. W przypadku modeli predykcyjnych na porządku dziennym jest oczekiwanie, że model lub metoda będzie analizowana z perspektywy odpowiedzialnego użycia oraz etyki użycia (ang. responsible machine learning). W rozprawie w rozdziale 4.5 metoda DiviK uzyskuje dobre wyniki ilościowe w odniesieniu do innych rozwiązań, ale czy taka ewaluacja jest wystarczająca? Dla rozwiązań trenowanych bez nadzoru, czyli algorytmów klastrowania, ta literatura nie jest jeszcze tak bogata jak w przypadku modeli nadzorowanych. Ale podczas obrony rozprawy oczekiwałbym szerszej dyskusji ryzyk stojących za zaproponowanymi algorytmami, metod ich głębszej walidacji i nadzoru np. w odniesieniu do nowych proponowanych regulacji takich jak Ustawa o SI (ang. AI act).

Przechodząc do uwag szczegółowych, pragnę zwrócić autorowi uwagę na kilka usterek lub braków, które utrudniają zrozumienie lub docenienie zaproponowanego rozwiązania.

- Nowe metody zaproponowane przez autora powinny być wyraźniej wyróżnione i opisane. Przykładowo na stronie 58 jest informacja o "Sampled GAP Statistic". Jak rozumiem to jest nowa oryginalna propozycja autora ale zabrakło mi jej bardziej formalnego opisu.
- Prezentując zastosowania algorytmu DiviK należy wprost wskazać kto i gdzie użył opisywanej metody. Obecnie w pracy znajdują się lakoniczne opisy typu "Another researcher uses DiviK to conduct quality control of the obtained Mass Spectrometry Imaging dataset" choć znacznie lepiej byłoby wskazać artykuł będący wynikiem tego użycia.

- Wprowadzając czytelnika pracy do tematyki autor stosuje seryjne odwołania do literatury. Na połowie strony 7 znaleźć można 20 pozycji literaturowych wrzucanych grupami bez dyskusji co jest w której pracy. Dodatkowo cytowane referencje nie są zbyt zróżnicowane jeżeli chodzi o zespoły, które się daną tematyką zajmują. Odnosząc się do zastosowań MSI w onkologii autor pisze "MSI promises answers to a range of questions regarding tumor molecular profiling [132, 94, 71]" wskazując tylko prace, których sam jest współautorem. Choć w literaturze znaleźć można też inne prace z tej tematyki i właśnie w tym miejscu należałoby pokazać swoje obycie z literaturą. Podobnie w zdaniu "Numerous filtering methods are applicable for high-throughput biological data [126, 136, 103, 80, 81, 60, 133]" jeżeli się nic nie napisze o tych siedmiu pracach to trudno oczekiwać, że czytelnik znajdzie uzasadnienie w umieszczeniu ich wszystkich w takiej ilości.
- Rozdział 3 choć opisuje metody analizy danych jest pozbawiony jakichkolwiek ilustracji, diagramów czy wzorów. Trudno się czyta taki lity blok tekstu, co więcej nie jest dobre opisywanie prozą wzorów matematycznych czy algorytmów. Powinny one być zapisywane w bardziej ustrukturyzowanej postaci. Np. niewiele wnoszą wyliczenia ze strony 20, można by je przestawić znacznie czytelniej np. tabelą.
- Z nieznanego mi powodu niektóre akapity są rozdzielone pustymi liniami (np. na stronie 24) a inne nie (np na stronie 25).
- Autor nie jest zbyt konsekwentny w sposobie umieszczania referencji. Czasem są obok nazwy metody, a czasem zamiast nazwy metody. Czasem obok a czasem zamiast nazwisk autorów. Ale najbardziej rażące jest zaczynanie całego akapitu od referencji jak to ma miejsce np. na stronie 26 w linijce "[65] provides a user guide to MSI data analysis".
- Wzór nie powinien rozpoczynać zdania jak to ma miejsce dwukrotnie na stronie 61.

- Przy odnoszeniu się do mniej znanych rozwiązań warto byłoby w rozprawie umieścić ich opis. Zdanie "Therefore, they suggested using an edge-preserving Chambolle algorithm [32] with a Grasmair modification [53]" jest mało czytelne jeżeli nie wiadomo jakie właściwości ma algorytm Chambolle'a i co daje korekta Grasmair'a.
- W rozprawie pojawiają się tezy, którym brakuje dowodu w postaci referencji lub argumentów przytoczonych przez autora. Przykładem jest np. zdanie "Surprisingly, most of the known approaches to unsupervised knowledge discovery in big -omics data still cover either feature engineering or observation grouping". Jestem gotów się z nim zgodzić, ale chciałbym zobaczyć argumenty autora rozprawy. Podobnie zdanie "An often simplification is to assume that all ions were loaded with a unit charge", jest prawdopodobnie prawdziwe, ale oczekiwałbym argumentów, najlepiej w postaci literatury.
- W rozprawie pojawiają się odwołania do nazw algorytmów ale bez referencji. W przypadku popularnych nazw to utrudnia zrozumienie o jaki algorytm autorowi chodzi. Przykładowo na stronie 17 pojawia się Top Hat i imzML ale bez referencji. Podobna sytuacja ma miejsce na stronie 21 i dotyczy algorytmu DBSCAN.
- Literatura ma kilka wad redakcyjnych. Obecne są w niej cytowania preprintów (np. Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426, 2018) choć należy tego unikać preferując cytowanie artykułów w recenzowanych czasopismach. Praca o algorytmie Umap jest opublikowana np w czasopiśmie Journal of Open Source Software (https://joss.theoj.org/papers/10.21105/joss.00861). Nazwy czasopism raz są pisane z użyciem dużej litery tylko na początku (np. The annals of statistics) czasem z dużymi literami w każdym wyrazie (np. Biochimica Biophysica Acta (BBA)-Proteins and Proteomics) a czasem tylko skrótem (np. PMLR). Nazwy własne są też niepoprawnie zapisywane jak np. MRI w 'Mri-compati-

ble pipeline for three-dimensional maldi imaging mass spectrometry using paxgene fixation'.

Powyższe krytyczne uwagi nie zmieniają mojej oceny realizacji głównego celu pracy, jakim było zaproponowanie i przetestowanie metody DiviK. Ten cel udało się autorowi zrealizować zarówno od strony metodologicznej jak i aplikacyjnej. Powstało oprogramowanie i dwie ilustracje użycia tego rozwiązania do prawdziwych zbiorów danych. Biorąc pod uwagę, że ta metoda jest również opublikowana w recenzowanym czasopiśmie stwierdzam, że rozprawa mgr. inż. Grzegorza Mrukwy spełnia warunki ustawowe stawiane rozprawom doktorskim i wnioskuję o dopuszczenie rozprawy doktorskiej do publicznej obrony.

Przemysław Biecek