Wpł. RDITT
dnia 04.11.2022
M. Skow

Dr hab. inż Zbigniew Świder, prof. PRz Rzeszów,
Katedra Informatyki i Automatyki
Politechnika Rzeszowska im. Ignacego Łukasiewicza
al. Powstańców Warszawy 12
35-959 Rzeszów

2.11.2022

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
Promotor pomocniczy: dr inż. Michał Marczyk
Dziedzina: nauki techniczne
Dyscyplina: informatyka techniczna i telekomunikacja

Niniejsza recenzja została przygotowana na zlecenie Rady Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Śląskiej.

1. Cel i zakres rozprawy

Rozprawa doktorska mgr inż. Grzegorza Mrukwy dotyczy technik klastrowania wysokoprzepustowych danych *omicznych* w zakresie eksperymentalnej biologii molekularnej oraz wspierania badań naukowych w dziedzinie biologii nowotworów. W rozprawie skupiono się na przedstawieniu stosowanych metod i propozycjach autorskich algorytmów dla przetwarzania danych *omicznych*.

Postawiono trzy główne cele badawcze:

1. Zastosowanie metodologii krokowej do grupowania danych *omicznych*, co pozwala na uzyskanie wyników porównywalnych z istniejącymi najnowocześniejszymi metodami jednoetapowymi.
2. Połączenie wysoce skalowalnych, ale prostych metod zapewniających dobrą skalowalność liczby obserwacji bez znaczącej utraty poziomu szczegółowości, co jest istotnym aspektem analizy danych *omicznych*, gdyż objętość zbioru danych znacząco rośnie zarówno pod względem liczby wymiarów, jak i liczby obserwacji.
3. Zbudowanie elastycznej metodologii analitycznej danych *omicznych*, łatwej do aktualizacji za pomocą najnowszych metod w celu zwiększenia wydajności, np. z wykorzystaniem głębokich sieci neuronowych.

W pracy dużą część zajmuje opis oraz modyfikacje klasycznych metod, których to odpowiednie połączenie pozwoliło na osiągnięcie bardziej kompleksowego wyniku, niż oryginalne podejścia spotykane w literaturze. W szczególności, kalibracja zwiększyła czułość metod do rozpoznawania niuansów biologicznych w silnie wielowymiarowej przestrzeni cech, natomiast automatyzacja

wyeliminowała potrzebę ręcznego wyszukiwania hiperparametrów i pozwoliła na wybrania tego optymalnego. Proponowane metody zostały porównywane z istniejącymi na dwóch różnych zestawach danych obrazowania metodą spektrometrii mas, obejmujących bardzo szczegółową próbkę 2D całej tkanki oraz dane 3D o wysokiej przepustowości.

2. Struktura i zawartość rozprawy

Recenzowana praca doktorska obejmuje formalnie 4 główne rozdziały, poprzedzone wstępem oraz zakończone podsumowaniem. Zasadnicza część rozprawy liczy łącznie 111 stron oraz dodatkowo zawiera bibliografię liczącą 140 pozycji.

Praca rozpoczyna się wstępem, w którym przedstawiono motywację i główne tezy pracy. Autor proponuje metodologię, która opiera się na wiedzy ekspertów w dziedzinie przetwarzania oraz cechach danych *omicznych*. Zastosowane podejście etapowe jest rdzeniem proponowanej metody, a klasyczny algorytm K-Means został dostosowany do grupowania danych MSI i połączony z metodą wyboru cech opartą na GMM.

W rozdziale 2 przedstawiono źródła danych wykorzystanych w tej pracy do celów porównawczych oraz opisano, w jaki sposób informacje biologiczne są pozyskiwane w formie cyfrowej. Omówiono także wpływ wybranej procedury akwizycji na jakość danych oraz proces wstępnego przetwarzania wymagany do rozwiązania występujących problemów.

W rozdziale 3 zaprezentowano obecny stan wiedzy w zakresie grupowania danych *omicznych*, w szczególności zwracając uwagę na zasadę ich działania, mocne strony, obszary ich zastosowania oraz znane ograniczenia.

W rozdziale 4 zaproponowano odporną i skalowalną metodę grupowania danych *omicznych*, zwaną Divisive Intelligent K-Means (DiviK). Wyjaśniono, w jaki sposób jest ona skalibrowana oraz w jaki sposób można liczbowo ocenić znaczenie biologiczne uzyskanych wyników. Na koniec rozdziału sprawdzono proponowany algorytm DiviK na zestawach danych przedstawionych wcześniej w rozdziale 2 oraz omówiono otrzymane rezultaty.

W kolejnym rozdziale opisano kolejny krok w rozwoju algorytmu DiviK, a więc grupowanie z podziałem za pomocą autoenkoderów wariacyjnych oraz połączenie DiviK z najnowszymi osiągnięciami głębokiego uczenia się, w celu ominięcia ograniczeń obu metod. Na końcu rozdziału zamieszczono otrzymane wyniki oceny algorytmu DiVAE ze zbiorem danych OSCC w porównaniu z najnowocześniejszymi metodami w tej dziedzinie.

W rozdziale 6 podsumowano otrzymane wyniki oraz wskazano, w jakich sytuacjach system Divisive Intelligent K-Means może być przydatny i z powodzeniem zastosowany. Podsumowanie jest poparte listą przykładów, w których DiviK został już z powodzeniem zaimplementowany.

3. Najważniejsze osiągnięcia rozprawy

Biorąc pod uwagę zawartość pracy oraz pozytywna ocenę jej zawartości merytorycznej, za główne osiągnięcia Autora należy uznać zaproponowanie metodologii opartej na doświadczeniach wielu ekspertów w dziedzinie przetwarzania danych oraz cechach danych *omicznych* omówionych w rozdziale 3. Zastosowane podejście etapowe jest rdzeniem proponowanej metodologii. Klasyczny algorytm K-Means został dostosowany do grupowania danych MSI i połączony z metodą wyboru cech opartą na GMM. W celu weryfikacji, przeanalizowano jakość uzyskanych segmentacji oraz sprawdzono poprawność oceny liczbowej.

Najważniejszymi elementami rozprawy decydującymi o jej wartości naukowej i badawczej są:

1. Stworzenie metodologii do nienadzorowanego badania danych *omicznych*, uwzględniającej jednocześnie problem inżynierii cech i klastrowania, która to jest wystarczająco elastyczna, aby można było w prosty sposób wymieniać w niej poszczególne komponenty.
2. Dostosowanie klasycznego algorytmu K-Means do grupowania danych MSI i połączenie go z metodą wyboru cech opartą na GMM.
3. Przeprowadzenie obliczeń na dużą skalę za pomocą proponowanej metody DiviK, dla weryfikacji jej zdolności do przetwarzania wielkich zbiorów danych.
4. Koncepcja przeniesienia podejścia DiviK na metodę DiVAE, co pozwoli wykorzystać najnowsze techniki w tej dziedzinie, jak na przykład głębokie uczenie.
5. Propozycja algorytmu DiVAE oraz (dla zbioru danych OSCC) porównanie wyników z najnowocześniejszymi metodami w tej dziedzinie.
6. Utworzenie repozytorium przeznaczonego do przeprowadzenia wstępnej eksploracji zbioru danych *omicznych* i ich dalszej analizy za pomocą metody Divisive Intelligent K-Means.

Należy zauważyć, że Autor podjął się realizacji bardzo ciekawego oraz istotnego z punktu widzenia praktycznych zastosowań tematu badawczego. Poszczególne wyniki badań Autora zostały opublikowane w kilku współautorskich pracach w języku angielskim, co świadczy pozytywnie o dużej wiedzy Autora rozprawy w zakresie poruszanej tematyki badawczej, popartej również doświadczeniem praktycznym, w tym również biegłości w programowaniu w języku Python.

4. Poprawność pracy i uwagi krytyczne

Poprawność treści rozprawy nie wzbudza zastrzeżeń, a stwierdzenia w niej zawarte wydają się być w pełni godne zaufania, co wynika w szczególności z przedstawionych podstaw teoretycznych popartych wynikami przeprowadzonych badań eksperymentalnych.

Jednocześnie Autor nie ustrzegł się pewnych drobnych niedociągnięć, a wśród uwag o charakterze krytycznym, a po trosze i dyskusyjnym, można wymienić:

1. Interesującym byłoby potwierdzenie, czy wyniki umieszczone w tabeli 4.4 odnośnie miar jakości każdego algorytmu (w pracy przetestowane dla danych OSCC) przeniosą się na dane z innych źródeł i również wykażą przewagę algorytmu DiviK. Czy ta przewaga algorytmu DiviK dotyczy tylko tych danych (i o podobnym charakterze), czy też można ogólnie stwierdzić, że algorytm ten daje lepsze wyniki dla większości danych medycznych i zalecane jest jego stosowanie?
2. Podobnie - czy wyniki uzyskane dla algorytmu DiVAE dla danych OSCC oraz zbudowanej sieci neuronowej (tabela 5.2) można bezpośrednio (bez większych modyfikacji) przenieść na inne bazy danych i uzyskamy znowu najlepsze wyniki w stosunku do klasycznych algorytmów?
3. Dla klasycznego algorytmu K-Means wymagany jest dedykowany etap kalibracji. Czy w proponowanej metodyce (lub w przyszłości) można to zautomatyzować, aby nie trzeba było „dostrajać ręcznie" algorytmu do nowej bazy danych medycznych?
4. Drobne uwagi szczegółowe (najważniejsze)

- Str. 10, 11, ... (i dalsze) – brak jednostek i opisów na osiach wykresów. (np. rys. 2.3, 2.4, ...)
- Str. 71 – niespójny rysunek (inna czcionka i rozmiar w porównaniu z innymi rysunkami)
- Str. 85, 91 – niespójne tabele, np. w tab. 4.5 są linie poziome, a w tab. 4.8 nie ma.

5. Podsumowanie

Przytoczone wyżej uwagi dyskusyjne nie umniejszają zasług Autora ani nie kwestionują przedstawionych osiągnięć, a opisywana w pracy problematyka dotyczy aktualnych i interesujących zagadnień naukowych. Recenzowana praca zasługuje na wysoką ocenę merytoryczną i wnosi istotny oraz oryginalny wkład w dziedzinę informatyki. Postawione cele i zadania pracy zostały w pełni zrealizowane, a jej tematyka dobrze wpisuje się we współczesny nurt badań w tym zakresie.

Syntetyczna o cena rozprawy:

A. Czy rozprawa zawiera oryginalne rozwiązanie problemu naukowego? Zdecydowanie TAK

B. Czy kandydat posiada ogólna wiedzę teoretyczną w dyscyplinie? Zdecydowanie TAK

C. Czy posiada umiejętność samodzielnego prowadzenia pracy naukowej? Zdecydowanie TAK

Stwierdzam zatem z pełnym przekonaniem, że opiniowana rozprawa Pana mgr inż. Grzegorza Mrukwy pt. „Clustering techniques of high-throughput big -omics data" zawiera samodzielne rozwiązanie ważnego i istotnego problemu naukowego, jednocześnie spełniając wszystkie wymagania przewidziane dla rozpraw doktorskich w aktualnie obowiązującej Ustawie o Tytule Naukowym i Stopniach Naukowych.

W związku z tym stawiam wniosek o **dopuszczenie rozprawy doktorskiej do publicznej obrony**.

Dr hab. inż. Zbigniew Świder