Dr hab. inż. Michał Kawulok, prof. PŚ
Politechnika Śląska
Wydział Automatyki, Elektroniki i Informatyki
ul. Akademicka 16, 44-100 Gliwice

Gliwice, 29 września 2023

Recenzja

rozprawy doktorskiej pt.
„Ewolucyjna agregacja rang w systemach rekomendacyjnych"
Pana mgr. Michała Bałchanowskiego

1 Podstawa wykonania recenzji

Rozprawa doktorska Pana magistra Michała Bałchanowskiego zatytułowana „Ewolucyjna agregacja rang w systemach rekomendacyjnych" została opracowana w roku 2023 na Wydziale Nauk Ścisłych i Technicznych Uniwersytetu Śląskiego, a jej promotorem jest Pani prof. dr hab. Urszula Boryczka. Wykonanie recenzji zostało zlecone dnia 26 lipca 2023 r. przez Dziekana Wydziału Nauk Ścisłych i Technicznych, Panią prof. dr hab. Danutę Struż, na podstawie decyzji Rady Naukowej Instytutu Informatyki Uniwersytetu Śląskiego w Katowicach z dnia 17 lipca 2023 r.

2 Tematyka rozprawy i ocena jej aktualności

Recenzowana rozprawa dotyczy zagadnień związanych z systemami rekomendacji, a w szczególności agregacji wyników generowanych przez różne metody służące do tworzenia list rekomendacyjnych. Zadanie rekomendowania produktów odpowiadających potrzebom i preferencjom konkretnych użytkowników lub ich grup jest znanym i intensywnie badanym obszarem stosowania sztucznej inteligencji. Tworzone rozwiązania algorytmiczne są systematycznie wdrażane w różnych systemach wyszukiwania informacji – są one powszechnie wykorzystywane w praktyce, przy czym wciąż trafność rekomendacji często rozmija się z oczekiwaniami użytkowników. Biorąc to pod uwagę należy stwierdzić, że podjęta tematyka badawcza jest aktualna, a opracowywane w tym zakresie rozwiązania odpowiadają na konkretne potrzeby rynkowe, przez co posiadają wysoki potencjał wdrożeniowy.

Oryginalnym osiągnięciem Doktoranta jest opracowanie ewolucyjnego algorytmu ewolucji różnicowej służącego do agregacji rang w systemach rekomendacyjnych. Warto w tym miejscu podkreślić, że tworzenie rozwiązań polegających na łączeniu wyników różnych metod stanowi intensywnie eksplorowany obszar badawczy, przede wszystkim w zakresie technik uczenia zespołowego realizujących zadania klasyfikacji danych. Metaheurystyki, a w szczególności metody ewolucyjne są powszechnie stosowane w zadaniach dotyczących klasyfikacji, natomiast możliwości ich wykorzystania do zadania agregacji rang są wyraźnie słabiej przebadane. Stwierdzam zatem, że Autor rozprawy prawidłowo zidentyfikował lukę badawczą i we właściwy sposób ukierunkował swoje prace badawcze.

3 Zawartość rozprawy i ocena jej układu

Rozprawa napisana została w języku polskim, składa się z 10 rozdziałów, słownika symboli, dodatku zawierającego rozszerzone wyniki badań eksperymentalnych, bibliografii oraz spisów: rysunków, tabel i algorytmów. Łącznie praca obejmuje 132 strony maszynopisu. Umieszczone na wstępie podziękowania oraz spis treści zajmują łącznie cztery strony, numerowane oddzielnie.

Rozdział 1 zawiera wprowadzenie do tematyki systemów rekomendacyjnych wraz z omówieniem szeregu ich praktycznych zastosowań. Autor przedstawił następnie motywację do badań nad agregacją rankingów, a także dobrze zilustrował problem za pomocą odpowiednich przykładów. W dalszej części rozdziału zidentyfikował lukę badawczą oraz postawił tezę rozprawy i w punktach przedstawił jej zasadnicze cele.

W rozdziale 2 znajduje się opis stanu wiedzy w zakresie systemów rekomendacyjnych. Na początku zdefiniowane zostały pojęcia, którymi Autor posługuje się w dalszej części pracy, pojawia się także formalna definicja systemu rekomendacyjnego oraz omówienie zastosowań komercyjnych. Kolejno zaprezentowana została taksonomia metod służących do tworzenia rankingów, przedstawione zostały główne wyzwania związane z dalszym rozwojem tych technik, a także zostały tam omówione wybrane algorytmy rekomendacyjne.

Rozdział 3 jest poświęcony problemowi oceny jakości generowanych rankingów. Omówione zostały znane z literatury metryki, w szczególności stosowane dalej w pracy (m.in. do obliczania wartości funkcji przystosowania) średnia precyzja oraz znormalizowany zdyskontowany skumulowany zysk. W rozdziale 4 Autor przedstawił wprowadzenie do metod ewolucyjnych, omówił ich zastosowania w systemach rekomendacyjnych oraz opisał algorytm ewolucji różnicowej. W kolejnym rozdziale przedstawione są techniki stosowane do łączenia rankingów generowanych przez poszczególne metody bazowe, w tym opisane są techniki wybrane przez Autora do porównań. W rozdziale 6 opisano zastosowania technik uczenia maszynowego do generowania rankingów.

Rozdział 7 przedstawia opracowany algorytm służący do łączenia wielu rankingów z wykorzystaniem ewolucji różnicowej. Ponadto Autor przedstawił istotną modyfikację algorytmu polegającą na uwzględnieniu rankingów użytkowników o charakterystyce zbliżonej do tego, którego ranking jest aktualnie tworzony.

Kolejne dwa rozdziały opisują kolejno środowisko badawcze wraz ze zbiorem danych wykorzystanym do walidacji eksperymentalnej (rozdział 8) oraz prezentację i omówienie uzyskanych wyników (rozdział 9). Rozprawę kończy rozdział podsumowujący uzyskane osiągnięcia i opisujący możliwe kierunki dalszych prac. Bibliografia składa się ze 193 pozycji, natomiast Dodatek A zawiera sześć rysunków przedstawiających szczegółowe wyniki przeprowadzonych prac eksperymentalnych.

Układ pracy oceniam jako w pełni poprawny. Autor zdecydował się podzielić pracę na wiele krótkich rozdziałów – można się zastanawiać, czy praca nie skorzystałaby na połączeniu ze sobą niektórych rozdziałów (przykładowo rozdział 6 składa się z zaledwie pięciu stron i mógłby zostać połączony z przeglądem zawartym w rozdziale 2), jednak praca w obecnej postaci jest wystarczająco przejrzysta, a jej struktura logicznie spójna.

4 Najważniejsze osiągnięcia rozprawy

Głównym osiągnięciem Autora rozprawy jest opracowanie nowego algorytmu agregacji rang (EAR) opartego o ewolucję różnicową, który służy do łączenia wielu rankingów generowanych przez różne systemy rekomendacyjne. Zaproponowane rozwiązanie jest oryginalne, a wyniki przeprowadzonych badań eksperymentalnych potwierdzają jego skuteczność i konkurencyjność względem innych uzna-

nych technik. Na uznanie zasługuje metodyka prac badawczych, obejmująca wykorzystanie testów statystycznych, a także dobrze zaplanowane eksperymenty, które pozwoliły na zrozumienie istotnych aspektów analizowanych algorytmów.

W rozprawie została postawiona teza o treści: „Zaproponowany algorytm ewolucyjnej agregacji rang poprawia jakość generowanej agregacji, w porównaniu z wybranymi metodami zaproponowanymi w literaturze." Co prawda tak postawiona teza pozwala na w zasadzie dowolną swobodę w doborze metod (kwestię doboru metod rozwijam w dalszej części recenzji), jednak przeprowadzone prace badawcze faktycznie pokazały, że opracowana technika stanowi interesujący wkład w badania nad systemami rekomendacyjnymi.

Warto podkreślić, że Doktorant jest świadomy pewnych ograniczeń zaproponowanej metody oraz zdaje sobie sprawę z tego, że przeprowadzone badania mogłyby zostać znacznie rozbudowane – dobitnie świadczy o tym dyskusja przedstawiona w rozdziale 10.3 wskazująca na możliwe ulepszenia metody oraz wyznaczająca kierunki dalszych prac. Zawarte tam wnioski stanowią ważną część rozprawy i w przeważającej większości zostały one sformułowane na podstawie badań opisanych w recenzowanej rozprawie. W szczególności interesujące byłoby opracowanie metody globalnej, która nie wymaga przeprowadzania optymalizacji dla każdego użytkownika niezależnie.

5 Uwagi krytyczne i dyskusyjne

Podczas lektury pracy nasunęły mi się następujące uwagi o charakterze dyskusyjnym, do których Doktorant powinien się odnieść podczas obrony:

1. Czy oszacowanie bazowe przedstawione wzorem (2.5) nie powinno stanowić średniej z trzech składników μ, b_{uA} i b_x? W obecnej postaci wartość takiego oszacowania może znacznie przekraczać liczbę 1.

2. Opracowany algorytm został poddany walidacji z wykorzystaniem relatywnie małego zbioru danych. W podsumowaniu Autor wskazał na konieczność przeprowadzenia testów z wykorzystaniem większych zbiorów danych, ale jednocześnie w rozdziale 9.4 wskazuje na czasochłonność algorytmu EAR. Prosiłbym Doktoranta, żeby przedstawił więcej szczegółów dotyczących czasu trwania optymalizacji oraz o omówienie podczas obrony kwestii związanych ze skalowalnością opracowanego systemu.

3. Opis metody bazującej na rozkładzie macierzy według wartości osobliwych w rozdziale 2.7.1 nie wyjaśnia jak dobierana jest wartość k podczas dekompozycji macierzy P (proces dekompozycji jest przedstawiony na rysunku 2.4). Jak jest ona ustalana?

4. Wątpliwości budzi definicja i opis precyzji oraz czułości (tak bym sugerował tłumaczyć termin *recall* zamiast zaproponowanego przez Autora terminu „zwrot") w systemach rekomendacji. Wyjaśnienie poniższych wątpliwości jest o tyle istotne, że precyzja została wykorzystana podczas obliczania wartości funkcji przystosowania.

    - Czy definiowanie czułości ma sens przy ustalonej odgórnie wartości K? Liczba przedmiotów relewantnych może być różna dla poszczególnych użytkowników, więc można sobie łatwo wyobrazić sytuację, że dla małych wartości K (mniejszych niż liczba elementów relewantnych) czułość zmierzona dla idealnie działającego systemu rekomendacyjnego byłaby dość niska. Chyba, że dla każdego ocenianego rankingu wartość K byłaby ustalana jako liczba przedmiotów w rankingu (który zawierałby wyłącznie przedmioty uznane

przez system jako relewantne). Co istotne, czułość nie jest wykorzystywana w dalszej części pracy.

- Czy we wzorze (3.3) długość listy ($|\tau_i^r@K|$) nie wynosi zawsze K? Chyba, że dopuszczalna jest sytuacja, że system rekomendujący zwróci listę krótszą niż K, ale nie zauważyłem, żeby w pracy taka możliwość była rozpatrywana.
- Podobne pytanie dotyczy średniej precyzji zadanej wzorem (3.5) i zilustrowanej na rysunku 3.2. Czy na przedstawionym przykładzie dla użytkownika u_1 system rekomendacyjny w rankingu uwzględnił przedmioty nierelewantne (na pozycjach 3, 4 i 5)? Czy nie powinniśmy w takim układzie oczekiwać, że ranking będzie zawierał wyłącznie dwa przedmioty x_3 i x_2? Elementy x_5, x_4 i x_1 można postrzegać jako elementy fałszywie pozytywne, które powinny obniżać precyzję rozumianą w analogiczny sposób jak przy ocenie jakości klasyfikacji.
- We wzorze (3.5) w mianowniku pojawia się $|r_i^r|$ – czy nie powinno się tam znaleźć $|\tau_i^r|$? Jeżeli tak jest, to czy przykład przedstawiony na rysunku 3.2 przyjmuje założenie, że dla obydwu użytkowników istnieją wyłącznie dwa przedmioty relewantne? Nie wybrzmiało to nigdzie w opisie, a gdyby było inaczej, to chyba mianownik miałby wartość większą od 2?
- Na rysunku 3.2 nie zostało określone znaczenie koloru zielonego i czerwonego – można się domyślać, że na zielono zaznaczono przedmioty relewantne dla użytkownika, a na czerwono – pozostałe (na tym założeniu oparłem wcześniejsze uwagi), jednak taki przykład nie powinien pozostawiać miejsca na domysły.

5. W przedstawionych badaniach zostały uwzględnione trzy warianty funkcji przystosowania. Mam tutaj dwa pytania. Czy fitness oparty wyłącznie o zbiór treningowy ($EAR_{DE_{F1}}$) nie będzie prowadził do wyboru metody, która najsilniej się do zbioru treningowego dostosowuje (czyli przeucza)? Uwzględnienie zbioru walidacyjnego przy obliczaniu wartości funkcji przystosowania wydaje się być dobrym kierunkiem, ale czy wzięto pod uwagę zastosowanie funkcji straty wykorzystującej wyłącznie zbiór walidacyjny?

6. Jak zostały dobrane wartości parametrów algorytmu ewolucji różnicowej (przedstawione w tabeli 8.9)?

7. Czy Doktorant brał pod uwagę możliwość strojenia parametrów algorytmów rekomendacyjnych z wykorzystaniem algorytmu DE? Optymalne wartości tych parametrów mogą się różnić dla poszczególnych użytkowników.

8. Jakim kryterium kierował się Doktorant wybierając metody agregujące, które zostały wykorzystane do porównań? Czy w jakikolwiek sposób wykorzystują one zbiór walidacyjny, czy też operują wyłącznie na podstawie zbioru treningowego? Szkoda, że opracowana metoda nie została porównana z żadną techniką ewolucyjną, na przykład jedną z tych, które zostały wymienione w tabeli 1.2.

9. Dla jakiego zbioru (treningowego / walidacyjnego / testowego) została przeprowadzona analiza wpływu parametru λ przedstawiona na rysunku 9.6?

10. W celu weryfikacji, czy różnice pomiędzy algorytmami są istotne statystycznie, Doktorant wykorzystał test Wilcoxona, jednak zabrakło informacji o tym jaka hipoteza zerowa była

weryfikowana i jaki został przyjęty poziom istotności testu. Wobec tego nie można stwierdzić, które z różnic przedstawionych w tabeli 9.10 są statystycznie istotne.

Zakładając, że hipoteza zerowa mówi o braku różnicy między metodami i przyjmując poziom istotności równy 0,05 (tak się można domyślać, nie wiadomo czy słusznie) można zauważyć, że różnice pomiędzy algorytmem $EAR_{DE_{F3}}$ a pozostałymi algorytmami agregującymi (z wyjątkiem CombMax i CombMin) nie są istotne statystycznie, zatem nie można powiedzieć, że ten algorytm był gorszy od algorytmów BordaFuse, CombMed i CombSum (a takie stwierdzenie pojawia się na str. 98). Średnie raportowane w tabeli 9.9 są co prawda niższe dla $EAR_{DE_{F3}}$, ale te różnice nie są statystycznie istotne. Co więcej, można zauważyć, że statystycznie istotne różnice względem wszystkich pozostałych metod pojawiają się wyłącznie w przypadku algorytmu $EAR_{DE_{F3(5)}}$ (co ważne, ten algorytm jest faktycznie lepszy od pozostałych metod agregujących), natomiast algorytm $EAR_{DE_{F3(10)}}$ co prawda uzyskał najwyższą uśrednioną skuteczność, ale jednak różnice pomiędzy nim a BordaFuse nie są istotne statystycznie ($p = 0,15$).

W związku z powyższymi rozważaniami prosiłbym Doktoranta o przedstawienie podczas obrony interpretacji uzyskanych wyników testu statystycznego i ewentualne zweryfikowanie wyciągniętych wniosków.

11. Zastanawia mnie dobór parametru k w klasyfikatorze kNN, który został wykorzystany do wyznaczania użytkowników podobnych. Czy Doktorant rozważał użycie dodatkowo (lub zamiast k) progowania wartości podobieństwa? Może się zdarzyć, że jakiś użytkownik będzie do tego stopnia osobliwy (różniący się zatem od innych użytkowników), że jego najbliżsi sąsiedzi będą i tak się mocno od niego różnić. Pytanie czy lepiej brać pod uwagę k najbardziej podobnych użytkowników, czy też wszystkich użytkowników o podobieństwie większym od jakiejś przyjętej wartości progowej.

6 Pozostałe drobne uwagi

Praca jest napisana w sposób jasny i przejrzysty, przy czym nie jest wolna od drobnych błędów językowych i redakcyjnych. Odnotowałem również pewne drobne nieścisłości, które są wypunktowane poniżej. Nie mają one charakteru dyskusyjnego, więc moim zdaniem nie ma potrzeby, żeby Doktorant się do nich odnosił podczas publicznej obrony.

- W bardzo wielu miejscach pracy przecinki są postawione błędnie (między innymi drugi przecinek w zdaniu „Chcemy mieć pewność, że nasz wybór, będzie właściwy." na str. 1). Natomiast w kilku miejscach ich brakuje (na przykład w zdaniu „...użytkowników dla których..." na str. 54). Takie przykłady (zwłaszcza nadmiarowych przecinków) można znaleźć po kilka prawie na każdej stronie.

- Odwołania do więcej niż jednej pozycji bibliograficznych powinny zawierać numery posortowane rosnąco (np. zamiast „[163, 54, 53]" powinno się pojawić „[53, 54, 163]").

- W pracy pojawiają się błędy fleksyjne, np. w pierwszym zdaniu na str. 2 powinno być „które" zamiast „który" (przedmioty mają zainteresować użytkownika, a nie ranking), w pierwszym zdaniu w rozdziale 2.6.1 powinno być „występujących" zamiast „występującym" itd.

- W Tabeli 1.1 pojawiają się liczby oznaczające jakość rekomendacji. W tekście pracy pojawia się informacja, że liczby te są z przedziału od 0 do 1, przy czym lepiej by było zamieścić taką informację również w podpisie tabeli.

- Na rysunku 1.1 (str. 3) pojawia się symbol τ oznaczający ranking, jednak symbol ten jest zdefiniowany dopiero w rozdziale 3.2.1 na str. 30. Podobnie jest w przypadku rysunku 1.2 – co prawda tam w podpisie pojawia się już odnośnik do tabeli zawierającej spis symboli, ale objaśnienie symboli powinno pojawić się także w tekście przy omawianiu rysunków.

- Praca zawiera niewielką liczbę literówek (m.in. na str. 5 w tabeli 1.2 w jednym miejscu powinno być Oliveira zamiast Olivera, a przy objaśnianiu tezy „średniej" zamiast „średnej," na str. 36 pojawia się „opracowli," „zwierający" zamiast „zawierający" (str. 63), w kilku miejscach powinno być 'ę' zamiast 'e' i odwrotnie).

- Definicja niektórych terminów jest powielona, na przykład pojęcie top-N rekomendacji (wraz z anglojęzycznym źródłem) pojawia się na str. 1 i 10.

- Wzory (2.10) i (2.11) w zasadzie powielają wzory (2.7) i (2.8) – jedyna różnica polega na tym, że odległość liczona jest pomiędzy użytkownikami, a nie pomiędzy przedmiotami.

- W rozdziale 2.7.2 Autor przedstawia za pomocą wzorów (2.15) – (2.17) własności porządku liniowego, jednak szkoda, że nie przytoczył nazw tych własności.

- W pracy pojawia się wiele niespójności natury typograficznej. Na przykład:
  - na str. 26 kryterium BPR-OPT pisane jest inną czcionką w tekście, a inną we wzorze,
  - na str. 31 również inną czcionką jest napisane MAP w tekście i we wzorze (3.6),
  - w rozdziale 4 kilka symboli inaczej jest napisanych w tekście, a inaczej w opisie pseudokodu na alg. 1 (np. F i $\mathbf{F}$).,
  - Problem klasy NP na str. 34 pisany jest kursywą (NP), a na str. 47 normalną czcionką (NP),
  - dotyczy to również nazwy algorytmu – czasami jest EAR, czasami EAR.

- Przykład przedstawiony na rysunku 3.1 jest moim zdaniem zbędny – w rozprawie doktorskiej nie ma potrzeby tłumaczyć na przykładzie czym różnią się miary MAE i RMSE.

- Brakuje odwołania do rysunku 4.1 w tekście pracy (siłą rzeczy jego zawartość nie jest też omówiona, nie zostało też w podpisie rysunku wyjaśnione znaczenie zastosowanych kolorów). Ponadto przedstawia on chyba jedynie pojedynczą iterację algorytmu ewolucji różnicowej, a nie jego pełny schemat jak sugeruje podpis rysunku.

- Nazwa łańcucha Markowa pochodzi od nazwiska (str. 45), wobec czego powinna być pisana z wielkiej litery.

- Skrót OC SVM (str. 50) nie został wyjaśniony.

- W rozdziale 9 przedstawiono wyniki uzyskane dla 25 losowo wybranych użytkowników (np. tabela 9.4), potem dla 300 wybranych użytkowników (np. tabela 9.7), a dopiero na końcu na pełnym zbiorze testowym. Nie zostało sprecyzowane, czy przedstawione wyniki cząstkowe zostały uzyskane dla podzbioru zbioru testowego.

- Uzyskane wyniki zostały w czytelny sposób zaprezentowane, przy czym analizując tabele 9.1 i 9.2 pojawia się pytanie, jak często metody agregujące uzyskują wynik lepszy niż najlepsza metoda bazowa (na przykład dla użytkownika nr 2 algorytm BPR uzyskuje 0,33, podczas gdy algorytmy agregujące uzyskują wartości gorsze (od 0,1 do 0,17). Taka analiza z pewnością pomogłaby lepiej zrozumieć zachowanie metod agregujących.

- Zamiast sformułowania „odstańcy" lepiej byłoby używać powszechnie przyjętego terminu „wartości odstające."

7 Wniosek końcowy

Przedstawiona przez mgr. Michała Bałchanowskiego rozprawa pt. „Ewolucyjna agregacja rang w systemach rekomendacyjnych" stanowi w mojej ocenie oryginalne rozwiązanie problemu naukowego, potwierdza ogólną wiedzę teoretyczną Doktoranta w dyscyplinie *informatyka* oraz wykazuje Jego umiejętność samodzielnego prowadzenia pracy naukowej. Przedstawione w recenzji uwagi krytyczne nie wpływają na moją jednoznacznie pozytywną ocenę rozprawy, tym samym spełnia ona wymagania określone w art. 187 ustawy z dnia 20 lipca 2018r. Prawo o szkolnictwie wyższym i nauce.

Biorąc powyższe pod uwagę, wnioskuję o dopuszczenie mgr. Michała Bałchanowskiego do dalszych kroków procedury uzyskania stopnia doktora nauk technicznych, w tym do publicznej obrony.

Signed by /
Podpisano przez:

Michał Jan Kawulok
Politechnika
Śląska

Date / Data:
2023-10-02 09:02