Prof. dr hab. Łukasz Wyrzykowski
Obserwatorium Astronomiczne
Uniwersytetu Warszawskiego
Al. Ujazdowskie 4
00-478 Warszawa
email: lw@astrouw.edu.pl

Warszawa, 26. lipca 2022r.

Recenzja pracy doktorskiej
mgra Artema Poliszczuka
Narodowe Centrum Badań Jądrowych
pt.
Zastosowanie metod uczenia maszynowego w astrofizyce i kosmologii

Tematyka rozprawy dotyczy badań nad aktywnymi galaktykami oraz ich fotometrycznej klasyfikacji i katalogowania za pomocą metod uczenia maszynowego. Podjęty temat naukowy jest istotny ze względu na konieczność tworzenia jak największych i najczystszych próbek galaktyk aktywnych na różnych przesunięciach ku czerwieni niezbędnych do badań nad ewolucją galaktyk oraz ich super-masywnych czarnych dziur. Zaprezentowane w pracy metody oraz jej wyniki mogą mieć szerokie aplikacje w różnych dziedzinach astrofizyki i kosmologii.

Praca doktorska napisana jest w języku polskim i w luźnym stopniu pokrywa się i częściowo wykracza poza zawartość dwóch opublikowanych prac doktoranta (z 2019 i 2021 roku) oraz kolejnej pracy będącej w przygotowaniu. Praca zawarta jest na 127 stronach, zawiera spis treści, spis tabel i rysunków, pięć głównych rozdziałów, bibliografię oraz dodatki, dzięki czemu spełnia wszelkie wymogi formalne dla prac doktorskich.

Poniżej omawiam kolejne rozdziały przedstawiając ich krótki opis oraz uwagi.

Wstęp

Wstępna część pracy doktorskiej jest bardzo krótka i w zasadzie jest rozszerzonym opisem zawartości całej pracy. Przedstawione są powody naukowe powstania pracy doktorskiej i zawartych w niej badań nad Aktywnymi Jądrami Galaktyk (AGN). Przede wszystkim przedstawiony jest argument za tworzeniem katalogów AGN-ów opartych o niekompletne dane, bez trudnych do uzyskania danych w średniej podczerwieni. Autor przedstawia też nowatorskie aspekty swoich badań, w szczególności zastosowania elementów logiki rozmytej w tworzeniu modelu klasyfikacyjnego. Pod koniec rozdziału Autor przedstawia skrótowy opis struktury dalszych rozdziałów.

Rozdział 2

Ten rozdział zawiera astrofizyczny opis przedmiotu badań: Aktywnych Jąder Galaktyk (AGN). Jest to w praktyce wstęp teoretyczny do zagadnienia AGN-ów. Autor przedstawia w nim w przekonujący sposób powód badań nad AGN-ami przedstawiony w całej pracy doktorskiej - niezbędne są kompletne i czyste próbki AGN-ów do badań astrofizycznych takich jak ewolucja galaktyk, ciemna materia czy wielkoskalowe struktury Wszechświata. Następnie Autor przechodzi do opisu źródeł emisji AGN-ów w wielu zakresach fal oraz do stosowanych w literaturze metod ich selekcji. Wskazane zostały obciążenia selekcji związane ze stosowaniem różnych zakresów długości fal. Rozdział zawiera bogate odwołania do literatury. W podrozdziale 2.2 zdecydowanie przydałoby się graficzne przedstawienie typowego SED AGNa wraz z zaznaczeniem omawianych zakresów fal, jak również zestawienie go z SED typowego dla galaktyk SFG. W następnym podrozdziale Autor kontynuuje wprowadzenie w zagadnienie badania AGN-ów z astrofizycznego punktu widzenia. Wskazuje na istotne aspekty badań AGN-ów, np. możliwość badania zależności między masą super-masywnej czarnej dziury a dyspersją prędkości gwiazd galaktyki-gospodarza.

Na stronie 5 niespójnie przedstawione są jednostki odległości obszaru BLR (pc vs promienie grawitacyjne). Na str. 7 pojawia się skrót WLQ, który nie został wcześniej zdefiniowany (określenie *weak line quasars* pojawia sie wyżej, ale nie jest tam zdefiniowany używany potem skrót). Na str. 17 niezrozumiała jest część zdania "… AGN-y selekcjonowane w zakresie podczerwieni są obecne w centralnych galaktykach małych halo.".

Rozdział 3

Ten rozdział zawiera opis danych obserwacyjnych użytych w pracy. Są to obserwacje pola północnego bieguna ekliptycznego wykonane przede wszystkim przez sondę AKARI w zakresie podczerwonym.

Wymienione zostały również obserwacje w innych zakresach fal elektromagnetycznych. Następnie Autor przechodzi do opisu procesu przygotowania próbek treningowej i generalizacyjnej, niezbędnych w dalszym procesie stosowania uczenia maszynowego. Zdefiniowane są podstawowe etykiety klas stosowane w całej pracy. Następnie Autor stosuje algorytm estymacji minimalnego wyznacznika kowariancji (MCD) na próbce generalizacyjnej - jest to jeden z nowatorskich pomysłów Autora zastosowany w tej pracy. Dzięki temu zabiegowi próbka generalizacyjna przyjmuje wielowymiarowy kształt zbliżony do próbki treningowej, co z kolei pozwala uniknąć problemu z niereprezentacyjnością próbki treningowej względem badanej próbki. Autor przedstawia cechy przygotowanych próbek w postaci tabeli z zakresami jasności w różnych pasmach. Rysunek 3.5 w szczególności świetnie obrazuje zaletę zastosowania algorytmu MCD.

Na str.20 pojawia się nazwa urządzenia JCMT bez rozwinięcia oraz wyraz submm, którego się nie stosuje bez rozwinięcia. Na str.24 pojawia się błędne odniesienie do rysunku 5.3 zamiast prawdopodobnie 3.1.

Rozdział 4.

Ten krótki rozdział zawiera szczegółowy matematyczny opis zastosowanych technik uczenia maszynowego (ML). Opisane są dokładnie różne rodzaje modeli i metod, które pojawią się w rozdziale 5. Zdefiniowane są też podstawowe pojęcia związane z używaniem metod ML, takie jak Precision czy Recall, jak również metody wykrywania obserwacji odstających.

W moim odczuciu ten rozdział powinien znaleźć się w znacznie wcześniejszej części pracy, najlepiej jako część ogólnego wstępu. Niektóre przedstawione pojęcia z dziedziny ML pojawiają się już we wcześniejszym rozdziale 3, na przykład dopiero w rozdziale 4 definiowane są próbki treningowe i generalizacyjne, które tworzone są już w rozdziale 3.

Rozdział 5.

Jest to główny rozdział całej pracy doktorskiej, w którym Autor opisuje przeprowadzoną procedurę tworzenia modelu klasyfikacyjnego i jego zastosowanie do zbudowania wynikowego katalogu AGN-ów. Rozdział jest bardzo obszerny i moim zdaniem z powodzeniem mógłby być rozdzielony na dwa osobne rozdziały a część jego elementów bardziej pasowałaby do wcześniejszych rozdziałów, np. opisane w rozdziale 5.3 szczegóły użytej logiki rozmytej lepiej pasowałyby do ogólnego wstępu, a procedura przygotowania danych algorytmem MCD do rozdziału 3.

Autor najpierw przygotowuje model wytrenowany na próbce treningowej, stosując różne algorytmy uczenia maszynowego. Testowanie algorytmów zostało przeprowadzone bardzo skrupulatnie stosując różne metody ważenia obiektów jak również stosując elementy logiki rozmytej. W celu optymalnego wykorzystania uczenia maszynowego i posiadając liczne pomiary w różnych filtrach, Autor zdefiniował cechy obiektów jako kolory, tj. kombinacje różnic między jasnościami w filtrach oraz przeprowadził proces selekcji cech najbardziej istotnych w oparciu o nowatorską metodę z wykorzystaniem statystyki KS. Następnie przetestowana została skuteczność algorytmów w wariantach z ważeniem opartym o odległość od środka klasy oraz o błędy pomiarowe. Dzięki zabiegowi ważenia można uniknąć problemów wynikających z niewielkiej liczby obiektów w próbce treningowej oraz innych obciążeń.

W podrozdziale 5.5 Autor przeprowadza ciekawą analizę otrzymanego katalogu AGN-ów, mianowicie poszukiwanie obiektów odstających, najpierw tych z błędnymi pomiarami photo-z, a następnie odstających w przestrzeni klas. Wykorzystany został algorytm Isolation Forest, który pozwolił skutecznie zidentyfikować obiekty, których fotometryczne przesunięcie ku czerwieni jest prawdopodobnie błędnie wyznaczone, co jest niezmiernie cenną informacją w wynikowym katalogu.

Wykrywanie obiektów odstających w oparciu o klasę było moim zdaniem najciekawszym wynikiem pracy doktorskiej. Taka analiza pozwoliła zidentyfikować AGN-y, które nie są typowe i różnią się znacząco od próbki treningowej, np. AGN-y typu II lub takie o dużym z. Autor przeprowadził tę analizę na różne sposoby stosując modele ML wytrenowane na galaktykach i AGN-ach. Ogromny potencjał tej części pracy nie został jednak moim zdaniem od końca wykorzystany. Zabrakło mi zatrzymania się nad przykładowymi obiektami odstającymi, dobrze widocznymi na rys. 5.18, oraz ich krótkiego omówienia w oparciu o dostępne obserwacje.

Rozdział 5 zakończony jest podrozdziałem Wnioski, który zawiera podsumowanie nie tylko rozdziału 5, ale też całej pracy doktorskiej, co nie jest dla mnie zrozumiałe, gdyż praktycznie takie samo podsumowanie pojawia się w rozdziale 6. Zastanowiło mnie również odwołanie do jednej z prac Autora (Poliszczuk i in. 2021), w której znalazła się szczegółowa analiza otrzymanego katalogu wynikowego i SED AGN-ów w oparciu o kod CIGALE. Wielka szkoda, że ten element analizy nie pojawił się w tym miejscu pracy doktorskiej, co znacznie by ją ubogaciło.

Na str.59 znajduje się błędne odwołanie do rysunku 3.1b (powinno być 3.1a). Na rysunku 5.19 zastosowano te same oznaczenia (zielony trójkąt) dla różnych typów obiektów.

Najważniejsze zalety pracy doktorskiej:

- Głównym wynikiem pracy jest otrzymanie nowego katalogu AGN-ów w ciekawym obszarze nieba, katalogu o zmaksymalizowanej czystości i kompletności, który może posłużyć do wielu zastosowań i dalszych badań.
- Katalog i uczenie maszynowe zostały wykorzystane do wskazania obiektów o błędnie zmierzonych fotometrycznych przesunięciach ku czerwieni oraz obiektów nietypowych.
- Autor wykazał się w pracy niezwykłą biegłością w dziedzinie uczenia maszynowego.
- Autor udowodnił doskonałą świadomość słabych stron zbudowanego systemu klasyfikacyjnego, np. pokazując, że nie udało się skutecznie odtworzyć selekcji opartej o dane rentgenowskie i podczerwone używając jedynie danych optycznych i podczerwonych.
- Przygotowana przez Autora metoda selekcji AGN-ów może znaleźć szerokie zastosowanie w astrofizyce i w dalszych badaniach tych obiektów, jak również może też zostać przeniesiona na grunt badań innych obiektów.

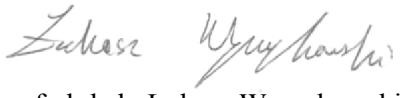
Poważniejsze uchybienia:

- Układ pracy nie jest moim zdaniem optymalny. Rozdz.2 zawiera wstęp na temat AGN-ów i ich obserwacji i klasyfikacji w dużych przeglądach. Podobnie wstępem jest też rozdział 4 o uczeniu maszynowym, natomiast w rozdziale 3 już pojawiają się operacje na danych w kontekście ML. Rozdz. 5 z powodzeniem mógłby być rozłożony na co najmniej dwa osobne rozdziały.
- W pracy brakuje pochylenia się nad wynikowym katalogiem i pokazania bardziej szczegółowych zastosowań. Poza dwoma zastosowaniami (błędne photo-z oraz obiekty odstające) nie zostały pokazane żadne przykłady ciekawych obiektów znalezionych w katalogu ani bardziej astrofizyczne jego zastosowania, np. do badań nad ewolucją galaktyk czy ciemnej materii. Można było na przykład pokazać, w jaki sposób otrzymany za pomocą ML katalog jest lepszy od innych analogicznych katalogów w jakimś konkretnym problemie kosmologicznym.
- Praca zawiera bardzo liczne błędy typograficzne, np. metoody (str.2), grifiki (str. 9), obserewacje (str.19), mu zamiast greckiej litery (str. 21, 27), Eq. zamiast Równanie (str.24), theta zamiast greckiej litery (str. 36), descet (str. 44), obietky (str.74) i kilka innych drobniejszych. Niespójne stosowanie przecinka dziesiętnego, wymiennie przecinek i kropka. Czytanie utrudniała duża odległość rysunków od opisującego je tekstu.

Podsumowanie:

Autor bardzo szczegółowo zbadał i przedstawił wszystkie mocne jak i słabe strony zastosowania uczenia maszynowego do danego problemu. Wykazał się biegłą znajomością problemu astrofizycznego jakim jest badanie AGN-ów, dzięki czemu wynikiem pracy jest cenny naukowo katalog AGN-ów. Mimo moich zastrzeżeń co do układu pracy jak i skromności przedstawionych zastosowań otrzymanych wyników, uważam, że praca doktorska jest nowatorska i wnosi znaczący wkład w badania naukowe.

W związku z tym uważam, że praca mgra Artema Poliszczuka spełnia wszystkie formalne i zwyczajowe normy pracy doktorskiej, dlatego wnoszę o jej dopuszczenie do publicznej obrony.



Warszawa, 26. lipca 2022r.

prof. dr hab. Łukasz Wyrzykowski