ul. Bartycka 18, 00-716 Warszawa
tel: (22) 841 00 41, (22) 3296 100
fax: (22) 841 00 46
email: camk@camk.edu.pl
http://www.camk.edu.pl

CENTRUM ASTRONOMICZNE IM. MIKOŁAJA KOPERNIKA PAN

Warszawa, 26 lipca 2022 roku

Dr hab. Michał Bejger
Centrum Astronomiczne
im. Mikołaja Kopernika PAN

Recenzja pracy doktorskiej mgr. Artema Poliszczuka pt. *„Zastosowanie metod uczenia maszynowego w astrofizyce i kosmologii"*

Recenzowana praca dotyczy dziedziny obserwacyjnej astronomii pozagalaktycznej i kosmologii. Mimo dość ogólnego tytułu zawiera wyniki konkretnej analizy (klasyfikacji) obserwacji obiektów pozagalaktycznych, galaktyk i aktywnych jąder galaktyk (AGN-ów) zarejestrowanych przez satelitę AKARI, który prowadził serię fotometrycznych pomiarów w podczerwieni w pasmach bliskiej (NIR), średniej (MIR) i dalekiej podczerwieni (FIR) w latach 2006-2011. Między innymi, AKARI wykonał głębokie obserwacje północnego bieguna ekliptycznego. Analiza danych z tego obszaru nieba stanowi podstawę pracy.

Temat i cel pracy jest dobrany trafnie w czasach coraz większej ilości dobrych danych obserwacyjnych, i potrzeb klasyfikacji obiektów w celu ich późniejszej dalszej analizy. Główną wynikiem pracy jest demonstracja zastosowania praktycznego metod: pokazuje kierunek rozwoju i dostarcza wyników oraz testów narzędzi potrzebnych do skonstruowania katalogu obiektów AGN na podstawie oceny ich cech obserwacyjnych w różnych filtrach optycznych i bliskiej podczerwieni, w celu otrzymania przewidywań dotyczących cech obserwacyjnych w średniej podczerwieni.

Praca opiera się o wyniki już opublikowane (artykuły Poliszczuk i in., 2019, 2021, cytowane w sumie 7 razy [NASA ADS]) i jest napisana w języku polskim. Na 101 stronach zawarto spis treści, rysunków i tablic, oraz 6 rozdziałów, bibliografię składającą się z 250 pozycji, i dwa dodatki. Rozdziały to 1 - wstęp, 2 - opis modelu AGN, 3 - opis danych wykorzystywanych w pracy, 4 - opis technik uczenia maszynowego wykorzystanych w pracy, 5 - opis modelu klasyfikującego i podsumowanie rezultatów, 6 - końcowe podsumowanie. Dodatki zawierają listę oprogramowania użytego podczas pracy oraz tabele z wynikami wartości metryk użytych do klasyfikacji danych, pochodzące z pracy Poliszczuk i in. (2021). Opis metod i wyników analizy to około połowa objętości pracy: 47 stron.

Z tekstu pracy nie wynika w oczywisty sposób, które wyniki z wyżej wymienionych prac są osobistym dokonaniem Autora, prócz stwierdzenia we wstępie o zmodyfikowaniu i wzbogaceniu wyników pracy Poliszczuk i in. (2021), dotyczących omówienia różnych

strategii logiki rozmytej oraz porównania ich wpływu na klasyfikacje z różnymi typami algorytmów klasyfikacji nadzorowanej, a także o dodatkowych badaniach nad nienadzorowanymi technikami wykrywania „obserwacji odstających" (ang. *outliers*).

Całość rozprawy jest napisana w sposób sprawny, a Autor wykazuje w niej znajomość astrofizyki badanych obiektów, szczegółów pozyskiwania i obróbki danych, a w szczególności implementacji różnych rodzajów metod uczenia maszynowego - klasyfikacji nadzorowanej opartej o regresję, metody typu maszyna wektorów nośnych (SVM) oraz drzewa decyzyjne. Autor przedstawia również dyskusję zastosowania logiki rozmytej w uczeniu nadzorowanym, wybór najlepszych cech obserwacyjnych (kolorów) do późniejszej klasyfikacji, oraz różne metryki oceny jakości klasyfikacji. Dodatkowo przedstawione są metody wykrywania danych typu *outlier* oraz wizualizacje danych wielowymiarowych poprzez nieliniową redukcję wymiarowości (metoda tSNE).

W trakcie czytania w wielu miejscach miałem uczucie niedosytu, bo tekst nie jest na tyle samowystarczalny, żeby dało się go czytać bez częstego sięgania do źródeł. Zdarzają się niewiele wnoszące do opisu lub błędne odnośniki: „patrz rozdział 3" (str. 22), „Rysunek 5.3" (str. 24). Nazwy metod pojawiają się w tekście wcześnie bez odnośnika do opisu (np. *maszyna wektorów nośnych* na str. 2, opis jest dopiero w rozdziale 4.1.2 na str. 36). W niektórych miejscach brakuje referencji lub choćby kilka zdań podstawowych wyjaśnień, np. w akapicie dotyczącym silnych klasyfikatorów (str. 41), definicji dywergencji Kullbacka-Leiblera (str. 44), metody PCA (str. 49), statystyki Kołomogorowa-Smirnowa (str. 49), metody tSNE i parametru *perplexity*, w szczególności na temat przyjmowanych przezeń wartości (str. 74). Zabrakło mi wstępnych podrozdziałów z testem/demonstracją metod na prostym syntetycznym zbiorze danych, które przydałyby się również do szacowania wiarygodności klasyfikacji realistycznych danych, i pokazałyby na przykład, że liczba danych treningowych jest wystarczająco duża (str. 23).

Inne otwarte pytania, które pojawiły się w trakcie czytania to:

- „Rosnąca odchylenie" i „rosnąca skośność w mniejszych skalach" (str. 17) - nie jest do końca jasne, do czego konkretnie odnoszą się te stwierdzenia,
- Czy *outlier* oznacza to samo, co *novelty*? (str. 43),
- Wybór cech (kolorów, np. N2-N4) jest przeprowadzony na podstawie wyników statystyki Kołogomorowa-Smirnowa (str. 49). Czy byłoby sensowne/wykonalne przeprowadzenie poszukiwania bardziej skomplikowanych kombinacji kolorów, np. dodatkową metodą uczenia maszynowego, w celu znalezienia cech optymalizujących konkretny cel klasyfikacji?
- Autor zauważa (str. 61), że „dane treningowe słabo reprezentują region zajmowany przez kandydatów na AGN-y". Czy w związku z tym jest możliwe przeformułowanie problemu/wybranie danych treningowych tak, by były lepiej dopasowane do fizyki problemu, czy też z jakiegoś powodu nie jest to możliwe?

Wizualizacja wyników jest najczęściej przeprowadzona porządnie, z niewielkimi wywołującymi zastanowienie wyjątkami, np. czemu rysunki w pracy po polsku są konsekwentnie opisywane po angielsku? Dlaczego wyniki przedstawione na rys. 5.5 (i innych podobnych), dotyczące *a priori* niezależnych metryk oceny dla różnych modeli są w formie punktów łączonych linią - sugeruje to jakiś związek między nimi i narzuca interpretację związaną z kolejnością rysowania. Na rys. 5.5A brak pomiaru jest wizualizowany wartością 0, natomiast na rys. 5.5B wartością 0,5. Czemu przy porównywaniu histogramów

mają one różną szerokość „binów” (np. rys. 3.4, 3.5, 5.17)? Szare i czarne kropki na rys. 5.18 są praktycznie nierozróżnialne, zwłaszcza w druku.

Zauważyłem także kilka nie do końca jasnych sformułowań, wynikających zapewne z tłumaczenia z języka angielskiego, m.in. „kąt nachylenia AGN” (str. 5), „torus . . . o częściowo zbitej strukturze” (str. 5), „modelowanie łączących się gospodarzy” (str. 7), „wysoce przesłonięte (obiekty)” (str. 10), „gładki (torus)” (str. 12), „spadek wartości koloru N2–N4” (str. 59).

Drobne uwagi redakcyjne. Praca nie jest pozbawiona literówek oraz problemów z LaTeX: „Zasosowana” (w Streszczeniu), brakujący czynnik 10^{38} w równaniu (2.1), „przesunieciac” (str. 13), „multiwavelenght” (str. 20 i 75), „3$sigma$” (str. 22), „wielowymiaroweej” (str. 26), $\vec{mu}$ (str. 27), problem z nawiasami w równaniu (4.5), „$theta_0$” (str. 35), „xi_i” (str. 36 i 37), „$\|theta_1\|$” (str. 36), $beta_m$ (str. 40), „gradient descet” (str. 44), „s” zamiast σ w podpisie pod rys. 5.16, brakujące znaki % w wartościach η_{Qi} dla modelu dopasowanego do danych AGN (str. 72), „obietky” (str. 74). Konsekwentnie stosowane są angielskie cudzysłowy (””), oraz kropka dziesiętna, zamiast „,” oraz przecinka dziesiętnego.

Wspomniane niewielkie niedociągnięcia nie wpływają jednak znacząco na ogólnie pozytywną ocenę pracy. Reasumując, przedłożona dysertacja spełnia ustawowe i zwyczajowe wymogi stawiane pracom doktorskim, dlatego wnioskuję o dopuszczenie mgr. Artema Poliszczuka do dalszych etapów przewodu doktorskiego.