Prof. dr hab. Przemysław Grzegorzewski

Wydział Matematyki i Nauk Informacyjnych
Politechnika Warszawska
ul. Koszykowa 75, 00-662 Warszawa

Warszawa, 27 grudnia 2021 r.

Recenzja rozprawy doktorskiej **mgra Pawła Piaseckiego** pt. **"Application of selected artificial intelligence methods to time series analysis"**

Niniejsza recenzja została napisana w odpowiedzi na pismo Dziekana Wydziału Matematyki i Informatyki Uniwersytetu Adama Mickiewicza, dra hab. Krzysztofa Dyczkowskiego, prof. UAM, skierowane do mnie w wyniku decyzji Rady Naukowej dyscyplin Matematyka i Informatyka, z dnia 7. lipca 2020 r., powołującej mnie na recenzenta rozprawy doktorskiej w przewodzie doktorskim mgra Pawła Piaseckiego.

1. Struktura i tematyka rozprawy

Jako rozprawę Doktorant przedłożył sześć artykułów naukowych, spośród których pięć (tj. [A1, A2, A4-A6]) zostało opublikowanych w latach 2018-2021, a jeden (tzn. [A3]) został przesłany do recenzji. Liczą one łącznie 101 stron i zostały poprzedzone pięciostronicowym wprowadzeniem. Dwie spośród opublikowanych prac ukazały się w czasopismach znajdujących się w wykazie MNiSW z 2019, a dokładniej: [A4] w *Communications in Statistics - Simulation and Computation* (40 p.) oraz [A5] w *Microchemical Journal* (70 p.), przy czym owa wyżej punktowana praca pochodzi z czasopisma nieprzypisanego do dyscypliny informatyka ani pokrewnej. Wszystkie sześć prac składających się na rozprawę doktorską mgra Piaseckiego są współautorskie. Do rozprawy dołączono oświadczenia określające wkład w prace poszczególnych współautorów (udziały Doktoranta wynoszą, odpowiednio, 80%, 80%, 60%, 80%, 30% i 80%).

Zwyczajowo tematyka rozprawy doktorskiej zostaje nakreślona – choćby zgrubnie – w tytule dysertacji. W rozważanym przypadku tytuł rozprawy, który w tłumaczeniu na język polski brzmi: **Zastosowanie wybranych metod sztucznej inteligencji do analizy szeregów czasowych**, ma niewiele wspólnego z zawartością prac ją tworzących. Jakkolwiek artykuły [A1-A4] dotyczą szeregów czasowych, to w pozostałych dwóch (tj. [A5-A6]) nie pojawia się nawet samo wyrażenie *time series*. Mylące jest również zawarte w tytule odniesienie do metod sztucznej inteligencji. W pracach [A1-A3] Doktorant bada przydatność rozmaitych miar odległości między szeregami czasowymi, co samo w sobie nie jest zagadnieniem z obszaru szeroko rozumianej sztucznej inteligencji. Praca [A4] dotyczy nowej miary zależności dla danych funkcjonalnych, co sytuuje ją raczej w obszarze nowoczesnej statystyki,

a z rezultatów raportowanych badań nie wynika by były one owocem technik wchodzących w zakres sztucznej inteligencji. I – paradoksalnie – jedynie w pracy [A6] dotykamy klasycznych technik sztucznej inteligencji, a mianowicie, sieci neuronowych.

Reasumując, wydaje się, iż tytuł niniejszej rozprawy, który by lepiej odpowiadał jej faktycznej zawartości, powinien brzmieć mniej więcej następująco: **Time series data mining – selected issues**.

2. Ocena rozprawy

2.1 Prace [A1-A3]

Pierwsza i zarazem najbardziej obszerna część rozprawy dotyczy miar odległości stosowanych w analizie szeregów czasowych. Wybór odpowiedniej odległości może bowiem mieć istotne znaczenie w kontekście choćby takich form wnioskowania jak klasyfikacja czy analiza skupień. W pracach [A1-A3] Doktorant badał empirycznie efektywność różnych odległości (pośród których nie wszystkie są odległościami w sensie ścisłym) na coraz liczniejszych zbiorach danych, przy czym w [A1] poddano badaniu 26 odległości na 34 zbiorach danych, w [A2] – 30 odległości na 47 zbiorach, natomiast w [A3] – 55 miar na 128 zbiorach danych.

Praca [A3] obejmuje nie tylko największy materiał doświadczalny, ale też jest najlepiej napisana spośród [A1-A3]. W szczególności, zadbano w niej o podanie definicji rozmaitych pojęć, którymi posługiwał się Doktorant, a które nie zawsze były wyjaśnione we wcześniejszych artykułach, jak i rozbudowano opis przeprowadzonego wnioskowania statystycznego służącego ocenie rozważanych odległości (por. rozdział 4 w [A3]). Na pochwałę zasługują:

a) rozmach przeprowadzonego badania, w którym uwzględniono wszystkie zbiory danych, które w danej chwili znajdowały się w znanej i cenionej bazie Uniwersytetu Kalifornijskiego (UCR *Time Series Classification Archive*) oraz być może wszystkie znane w danym momencie miary odległości między szeregami (przy czym niektóre z nich zostały po raz pierwszy zaimplementowane przez autorów omawianej pracy [A3]);

b) pomysłowe i po części nowatorskie metody przeprowadzonej analizy statystycznej i porównania klasyfikatorów szeregów czasowych (jak choćby grupowanie odległości, czy porównanie parami dla najlepszych metod w celu wykazania istotnych różnic);

c) wprowadzenie nowej kategorii miar odległości dla szeregów czasowych tworzonych z wypukłej kombinacji kilku odległości (tj. *combined distances*).

Przyznam, że praca [A3] bardzo mi się podoba – dotyczy ważnego zagadnienia, została przejrzyście i ciekawie napisana oraz pięknie zilustrowana. Gdybym był jej recenzentem wydawniczym rekomendowałbym przyjęcie jej do druku z uwagi na potencjalnie dużą przydatność dla praktyków zajmujących się analizą szeregów czasowych, jak i badaczy pracujących nad nowymi miarami odległości.

Oceniając łącznie prace [A1-A3] nietrudno zauważyć, że artykuły [A1] i [A2] są mniej dojrzałymi wersjami pracy [A3]. Patrząc z tej perspektywy szkoda, że poszerzając materiał doświadczalny Doktorant nie rozbudował w większym stopniu metodologii porównań miar odległości. Przykładowo, można by spróbować wyjść poza przyjęty do badania klasyfikator 1NN i sięgnąć po inne klasyfikatory, jak również ocenić stabilność wyników uzyskiwanych przy różnych kryteriach. Analizując podobień-

stwo szeregów czasowych warto byłoby również rozważyć nie tylko miary odległości (ang. *distance*), ale i nieco odmienne konstrukcje znane z literatury przedmiotu, jak „rozbieżność" (ang. *divergence*), „niepodobieństwo" (ang. *dissimilarity*), etc.

2.2 Praca [A4]

Praca [A4] dotyczy pewnego obszaru wnioskowania statystycznego, który od kilkunastu lat przyciąga uwagę wielu statystyków, a którym jest analiza danych funkcjonalnych. Ponieważ szeregi czasowe mogą być postrzegane jako dane funkcjonalne z czasem dyskretnym, a z drugiej strony, dane funkcjonalnych stanowią swoistą reprezentację szeregów czasowych, praca ta wiąże się z tematem recenzowanej rozprawy.

W artykule [A4] zaproponowano uogólnienie na przypadek danych funkcjonalnych współczynnika Prokrustesa (znanego również jako współczynnik RLS), używanego m.in. w statystycznej analizie kształtu. Wspomniana praca wpisuje się w cykl artykułów autorstwa grupy statystyków z UAM, poświęconych analizie korelacji wielowymiarowych danych funkcjonalnych. Jednakże prowadzone w niej rozważania wiodą tym razem do nieco bardziej złożonych formuł matematycznych (niż np. w przypadku współczynników funkcjonalnych ρV lub dCov), wymagających przeprowadzenia odpowiednich obliczeń numerycznych. W pracy zamieszczono także przykład ilustrujący zastosowanie wprowadzonego narzędzia na ciekawym zbiorze rzeczywistych danych dotyczących czaszek makaków orientalnych (*Macaca nemestrina*). Omówiono również wpływ doboru bazy (jak baza Fouriera, B-splajny) na wartość współczynnika.

Podsumowując, zaproponowane uogólnienie współczynnika Prokrustesa niewątpliwie wzbogaciło katalog narzędzi do oceny związku między danymi funkcjonalnymi. Pewnym mankamentem pracy [A4] jest brak pogłębionej analizy własności statystycznych zaproponowanego współczynnika oraz porównania z innymi miarami zależności.

2.3 Praca [A5]

Praca [A5] jest efektem współpracy Doktoranta z uczonymi z Uniwersytetu Przyrodniczego w Poznaniu, prowadzonej w ramach grantu NCN. W rzeczonym artykule mamy do czynienia z danymi dotyczącymi spektrometrii żywnościowej, przyjmującymi postać szeregów czasowych. Jednakże ich analiza stanowi wyłącznie tło do rozważań dotyczących różnych aspektów badania jakości jabłek, a udział Doktoranta, poświadczony podpisami, nie jest powiązany z metodologią badań ale z działalnością określoną jako „software, formal analysis" i obejmującą eksplorację i wstępne przetwarzanie danych oraz poszukiwanie najlepszego modelu regresji dla kilku parametrów chemicznych.

Być może, gdyby opisać dogłębnie przebieg badań prowadzonych na rzecz pracy [A5], udałoby się wydobyć i podkreślić jakieś wątki wskazujące na metodologiczny wkład doktoranta w analizę szeregów czasowych (jeśli takowy faktycznie miał miejsce). Jednakże w obecnym kształcie artykuł [A5] eksponuje wyłącznie cel, dla którego został napisany, tj. porównanie różnych technik spektroskopii optycznej i ich przydatności do oceny jakości soków jabłkowych.

2.4 Praca [A6]

W ostatniej z prac cyklu, przedłożonych jako rozprawa doktorska, tj. [A6], zaprezentowano pomysł budowy nowego klasyfikatora, nazwanego losową siecią neuronową (*Random Neural Network*), łączącego zalety lasów losowych i sieci neuronowych. Osiąganiu za jego pomocą dobrych wyników klasyfikacji ma sprzyjać połączenie trzech źródeł losowości: baggingu, losowego doboru cech oraz losowej modyfikacji architektury sieci.

Być może przyszłość potwierdzi intuicje autorów artykułu, jednakże – póki co – wyniki zawarte w omawianej pracy mają charakter wstępny, co stwierdza sam Doktorant w streszczeniu rozprawy.

2.5 Uwagi ogólne

Zgodnie z ustawą *Prawo o szkolnictwie wyższym i nauce*, Art. 187, p. 3: „Rozprawę doktorską może stanowić praca pisemna, w tym monografia naukowa, zbiór opublikowanych i powiązanych tematycznie artykułów naukowych, praca projektowa, konstrukcyjna, technologiczna, wdrożeniowa lub artystyczna, a także samodzielna i wyodrębniona część pracy zbiorowej". W rozważanym przypadku trudno mówić o silnym powiązaniu tematycznym artykułów, za wyjątkiem prac [A1-A3] (ewentualnie, można doszukiwać się pewnego powiązania pracy [A4] z [A1-A3]). Pozostałe prace, tzn. [A5] i [A6], nie są powiązane tematycznie ani ze sobą nawzajem, ani z wcześniejszymi pracami i choć ciekawe same w sobie, mają raczej charakter przyczynkowy. Oczywiście, są one świadectwem zaangażowania doktoranta w prace badawcze, ale celowość ich włączenia do rozprawy budzi wątpliwości.

Decydując się na rozprawę w formie zbioru powiązanych tematycznie artykułów naukowych należy być świadomym związanych z tym konsekwencji. W szczególności, rozprawa skomponowana z prac o zbyt podobnej tematyce zawiera – siłą rzeczy – liczne powtórzenia, a nieraz wręcz autocytowania w dosłownym brzmieniu, z czym mamy do czynienia w recenzowanej dysertacji (por. prace [A1-A3]). Z drugiej strony dołączenie do cyklu prac odbiegających zanadto od głównego nurtu rozważań grozi utratą zaleconego przez ustawodawcę powiązania tematycznego artykułów naukowych.

Dodatkowo, trudno dociec, dlaczego rozprawie nadano tytuł niezbyt odpowiadający jej zawartości. Wszak Doktorant z pewnością wie, jakiego typu narzędzia, metody i techniki wiążą się z pojęciem sztucznej inteligencji. W przypadku analizy szeregów czasowych są to przede wszystkim rozmaite podejścia związane z głębokim uczeniem i ogólnie pojmowanymi sieciami neuronowymi, z wnioskowaniem rozmytym, algorytmami ewolucyjnymi, systemami regułowymi oraz szeroką gamą metod hybrydowych. W recenzowanej rozprawie, niestety, nie padło na ten temat ani słowo. Zabrakło też odniesień do głównych nurtów badań nad szeregami czasowymi, uprawianymi w ramach sztucznej inteligencji, jak choćby budowanie reguł; posumowania lingwistyczne; agregacja, segmentacja i granulacja danych oraz zastosowanie owych technik w prognozowaniu; wykrywanie anomalii, wyjaśnianie modeli, analiza przyczynowości, etc.

3. Podsumowanie

Przedstawione wyżej refleksje prowadzą wg mnie do prostej konkluzji, iż rozprawa doktorska mgra Piaseckiego zyskałaby i to znacznie – gdyby nadano jej formę klasycznej, jednolitej, dysertacji, a nie zbioru artykułów.

Niemniej, mimo wielu krytycznych uwag, pragnę docenić wiedzę i szerokość horyzontów naukowych Doktoranta, wielość podjętych przezeń wyzwań badawczych i to, jak sobie z nimi poradził w praktyce oraz faktyczne osiągnięcia naukowe, a zwłaszcza te, które opisano w artykułach [A3] i [A4]. I te właśnie aspekty skłaniają mnie do ostatecznie pozytywnej oceny recenzowanej rozprawy.

Tym samym uznaję, iż **przedłożona rozprawa mgra Pawła Piaseckiego spełnia formalne i zwyczajowe wymagania stawiane pracom doktorskim z informatyki i wnoszę o dopuszczenie Doktoranta do dalszych etapów przewodu doktorskiego**.