POLITECHNIKA WARSZAWSKA
WYDZIAŁ ELEKTRONIKI I TECHNIK INFORMACYJNYCH
INSTYTUT INFORMATYKI
00-665 Warszawa, ul. Nowowiejska 15/19
tel. (22) 234 7432 NIP 525-000-58-34

Warszawa, 17 czerwca 2024 r.

dr hab. inż. Robert Nowak, prof. uczelni
Instytut Informatyki
Wydział Elektroniki i Technik Informacyjnych
Politechnika Warszawska
ul. Nowowiejska 15/19
00-665 Warszawa

Recenzja rozprawy doktorskiej mgr. inż. Łukasza Kupracza zatytułowanej „Metody predykcji w szeregach czasowych oparte na sztucznej inteligencji w wybranych procesach złożonych"

Recenzja powstała na prośbę Rady Naukowej Dyscypliny Informatyka Techniczna i Telekomunikacja Zachodniopomorskiego Uniwersytetu Technologicznego w Szczecinie, wyrażonej w piśmie z dn. 17.04.2024. Podstawą recenzji jest rozprawa doktorska licząca 256 stron, z kwietnia 2024 r. oraz dorobek publikacyjny Kandydata, uwzględniony w bazach Web of Science, Scopus i Google Scholar.

1 Tematyka badań

Rozprawa dotyczy nowych metod predykcji szeregów czasowych obrazujących zjawiska w systemach złożonych. Tytuł rozprawy odpowiada jej treści. Mgr inż. Łukasz Kupracz zaproponował nowe metody analizy tego typu danych na przykładzie analizy liczby zakażeń wirusem SARS-CoV-2 (COVID-19) oraz inflacji, dostarczając modele pozwalające na predykcję. Rozprawa opisuje autorskie implementacje przedstawionych metod w języku Python albo w środowisku Matlab.

Cel badawczy jest postawiony właściwie, jest on interesujący i istotny.

2 Główne wyniki rozprawy

Rozprawa liczy 256 stron, składa się z 5 rozdziałów, gdzie:

- rozdział 1; początkowe 24 strony pokazują znaczenie przewidywania inflacji i liczby zachorowań, później są opisane typowe metody analizy szeregów czasowych, zaś ostatnie 2 strony opisują problem badawczy, który jest przedmiotem rozprawy;
- rozdział 2 zawiera 56 stron i opisuje wybrane metody predykcji, które bada i modyfikuje mgr Kupracz;
- rozdział 3 liczący 73 strony opisuje badania metod predykcji zachorowań uwzględniając wpływ parametrów na wyniki;

- rozdział 4, liczący 13 stron, opisuje własne metody predykcji inflacji z ich badaniami;
- rozdział 5, 11 stron, zawiera podsumowanie, wnioski i dyskusję.

Na podstawie rozprawy nie można jednoznacznie stwierdzić, że zostało przedstawione oryginalne rozwiązanie problemu badawczego. Ponadto nie można jednoznacznie stwierdzić, czy Kandydat wykazuje umiejętność samodzielnego prowadzenia pracy naukowej w obszarze Nauk Technicznych w dyscyplinie Informatyka Techniczna i Telekomunikacja. Nie pokazano także, że metody są nowe, a wyniki wyniki badań są porównywalne z wynikami znanymi z literatury.

W rozprawie w sposób właściwy przeprowadzono analizę źródeł z literatury światowej, co świadczy o dostatecznej wiedzy Kandydata. Wnioski z przeglądu źródeł sformułowano w sposób jasny i przekonujący, pozwoliły one postawić opisane w pracy problemy badawcze, dotyczące metod predykcji szeregów czasowych dla systemów złożonych.

Przedłożona rozprawa doktorska ma właściwy cel badawczy, natomiast opis jest wadliwy, co nie pozwala jednoznacznie ocenić istotnych cech. Stwierdzam, że **rozprawa wymaga wprowadzenia poprawek i ponownego recenzowania**. Proponowane poprawki zamieściłem w punkcie 4.

3 Dorobek publikacyjny Kandydata

Osiągnięcia naukowe Kandydata mierzone publikacjami nie są wysokie, ale wystarczające. W bazach Web of Science lub Scopus lub Google Scholar znalazłem informacje o trzech publikacjach, których współautorem jest Kandydat. W żadnej nie jest autorem pierwszym (wiodącym).

- A comparative study of the multiple intelligence profiles of first-year IT students and employed graduates, A Wiliński, **Łukasz Kupracz**, Informatics in Education, 2020, IF=2.7;
- COVID-19: About the inequality of the territorial distribution of the number of deaths in the world and the indicator of the quality of epidemic management, A Wilinski, MK Arti, **Łukasz Kupracz**, IEEE Transactions on Computational Social Systems, 2022, IF=5;
- A Method of Selecting Computer Science Students for the IT Market Based on their Predispositions Resulting from Multiple Intelligence, Antoni Wilinski, Mariia Skulysh, Arti MK, Irena Bach-Dabrowska, Abayomi O Agbeyangi, Hina Zahra, Hubert Krason, Jolanta Dobska, **Łukasz Kupracz**, Informatics in Education, 2022, IF=2.7.

Ponieważ praca w czasopiśmie IEEE Transactions on Computational Social Systems jest wieloautorska i dotyczy przedstawianych metod, konieczne jest bardziej precyzyjne określenie wkładu Kandydata w rozwój omawianych w rozprawie metod.

4 Propozycja uzupełnienia rozprawy

Proponowane uzupełnienia i modyfikacje rozprawy są następujące:

1. szerszy opis własnych metod i wyraźne określenie indywidualnego wkładu,
2. porządkowanie opisu badań, porównanie wyników własnych z wynikami innych badaczy,
3. poprawa spójności rozprawy,
4. redukcja nieistotnych informacji,
5. poprawa redakcji technicznej.

4.1 Szerszy opis własnych metod, określenie indywidualnego wkładu

Cel szczegółowy badań jest opisany na str. 71 rozprawy. Jest nim opracowanie sposobów tworzenia modeli predykcyjnych, aby dla wybranego państwa i interwału czasowego, poprawnie prognozować parametry pandemii i inflację. Kandydat uzupełnia cel prac o zalecenie, aby błąd predykcji, w porównaniu z metodą referencyjną, był jak najmniejszy.

Celem rozprawy było opracowanie metod i modeli statystycznych do prognozowania. W wyniku badań powstały 3 metody predykcji szeregów czasowych zakażeń COVID-19, oraz 2 metody do predykcji szeregów czasowych inflacji. Metody te to:

- multiregresja z wykorzystaniem pseudo-inwersji;
- predykcja oparta na łańcuchach Markowa;
- predykcja oparta na podobieństwie do najbliższych krajów, bazująca na metodzie k najbliższych sąsiadów;
- regresja wielomianem;
- regresja z wygładzaniem.

Potrzebny jest systematyczny i poprawny opis tych metod. Proponuję wykorzystać pseudo-kod, schemat blokowy lub jakiś inny formalizm. Posługując się takim opisem metody proszę wyraźnie zaznaczyć własne elementy, rozszerzenia i modyfikacje. Proszę skupić się na opracowanych metodach i ich właściwościach.

W przekazanej do recenzji pracy taki opis każdej z metod jest długi i przeplata się z przykładem użycia. Nie wiadomo, które metody są nowe, a które istniejące. Nie wiadomo, które elementy są typowe, a które zmodyfikowane przez Kandydata. Czy wszystkie metody w rozdziale 2 są nowe i mają być oceniane jako wnoszące wkład w Dyscyplinę Informatyka Techniczna i Telekomunikacja? W rozdziale 2.1 jest opisana metoda multiregresji, w rozdziale 2.2 metoda oparta na łańcuchu Markowa, zaś w 2.3 metoda oparta o k-najbliższych sąsiadów (Kandydat nazywa ją N-najbliższych sąsiadów). Później w rozdziale 2.4 mamy 3 metody prognozy inflacji, ale z opisów są to metody istniejące, zmiany i rozszerzenia nie są wyraźnie zaznaczone. Z kolei rozdział 3 bada wpływ dodatkowy parametrów na metody z rozdziału 2.1, 2.2, 2.3, zaś rozdział 4 bada

algorytmy predykcji inflacji, przy czym przy badaniach powielany jest częściowo opis metody.

Proszę wyraźnie i jednoznacznie pokazać własne elementy i uzasadnić ich wprowadzenie.

4.2 Uporządkowanie i uzupełnienie opisu wyników

W pracy brak jest porównania wyników uzyskiwanych przez proponowane metody i proponowane modyfikacje z wynikami innych metod, znanymi z literatury. Wszystkie porównania dotyczą metod dostarczonych przez Kandydata. Podobnie badania wpływu parametrów czy wpływu uwzględnienia dodatkowych danych na wyniki dotyczą wyłącznie metod Kandydata.

Zarówno inflacja jak i parametry związane z pandemią COVID-19 są szeroko badane i istnieje wiele publikacji, część z nich Kandydat przytacza w rozprawie. Pożądane jest porównanie uzyskiwanych wyników do wyników opisywanych przez innych badaczy.

Kolejnym mankamentem obecnej rozprawy, który proponuję wyeliminować przy okazji poprawiania rozprawy, jest obecność wielu badań, wykresów, tabel, które nie są komentowane i nie są wyciągane wnioski. Przykładem jest Rys. 26. Po co go pokazano? Czy jest odwołanie w tekście?

W wielu miejscach wnioski są dyskusyjne. Lista jest długa, ale ponieważ wnioskuję o ponowną recenzję, to ten element krytyczny pominę pokazując przykład ze str. 221, gdzie napisano: „Wbudowane funkcje optymalizacyjne w regresji wygładzania splajnem okazały się dokładniejsze niż ręczna optymalizacja parametrem wygładzania p". Nie wiadomo, co to znaczy „ręczna optymalizacja", czy dobór parametrów na podstawie kilku prób, czy parametry domyślne.

Innym przykładem nieprecyzyjnego opisu jest badanie wydajności rozwiązania, cytując (str 219):

> Szybkość wykonywania obliczeń w predykcji jest atutem rozważanych metod – prawie wszystkie symulacje wykonywane są w czasie poniżej 5 sekund, najdłużej wykonywana jest predykcja oparta na podobieństwie n-sąsiadów i związana z tym analiza korelacji wzajemnej krajów. Obliczenia dla parametru n-sąsiadów n>9 i horyzontu predykcji hp=14 zajmuje około 4 minut. Szybkość wykonywanych prognoz oraz niskie zapotrzebowanie na moc obliczeniową umożliwia wdrożenie metod na sprzętowych rozwiązaniach, np. telemetrycznych, pojazdach autonomicznych oraz innych urządzeniach Internetu rzeczy.

Prosiłbym o podanie złożoności obliczeniowej (czasowej i pamięciowej) dla każdej z metod, a jeżeli mierzymy czas działania, to (1) dla jakiej paczki danych, (2) na jakim komputerze.

R

4.3 Dbałość o spójność pracy

Plan badań, cel pracy jest w rozdziale pierwszym (punkt 1.9), ale także w rozdziale 3.1 (plan prac badawczych). Co gorsze, cele są różne. Proszę umieścić ten element w jednym miejscu.

Opis istniejących metod pojawia się w rozdziale pierwszym: podrozdział 1.7 „Powszechnie stosowane metody predykcji rozprzestrzeniania się wirusów", strony 37-58, oraz podrozdział 1.8.1 „Modele inflacji i metody jej obliczania", ale też w rozdziale 2.

Celem badań (str. 71 rozprawy) jest znajdowanie metod, aby dla wybranego państwa i interwału czasowego dostarczać model zmian w pandemii i inflacji pozwalający na ich prognozowanie, oraz aby błąd w porównaniu z metodą referencyjną był jak najmniejszy. Dlaczego więc nie stosowano multiregresji z wykorzystaniem pseudo-inwersji, predykcji opartej na łańcuchach Markowa, czy predykcji opartej na podobieństwie do najbliższych krajów do predykcji inflacji? Dlaczego nie użyto regresja wielomianem, czy regresja z wygładzaniem do predykcji zachorowań w pandemii? Wszystkie pokazane metody można stosować do analizy szeregów czasowych, więc warto pokazać ich działanie na analizowanych zbiorach danych.

Separacja metod (osobne do danych pandemicznych, osobne do inflacji) oraz separacja obszarów zastosowań sprawia, że praca nie jest spójna, sprawia wrażenie poszukiwania rozwiązania dwu odrębnych problemów. Jeżeli Autor uznał, że opracowana przez niego metoda do predykcji zachorowań jest nieodpowiednia do predykcji inflacji, to proszę to stwierdzić i uzasadnić.

4.4 Redukcja nieistotnych informacji

Praca jest obszerna, występuje sporo powtórzeń. Główne źródło to powtarzanie argumentów o istotności predykcji wybranych szeregów czasowych, wielokrotne wprowadzanie tych samych pojęć, omawianie przykładów pokazując drobiazgowo kolejne etapy obliczeń i pisanie treści, która jest oczywista.

Praca została przypisana do dyscypliny Informatyka Techniczna i Telekomunikacja, więc proponuję dyskusję o istotnej roli inflacji i liczby zachorowań na COVID-19 ograniczyć do wstępu.

Przy wprowadzaniu pojęć proszę robić to raz i później się odwoływać. Przykład: kilkadziesiąt razy w pracy jest używany akronim MAPE (Mean Absolute Percentage Error). Na stronie 73, jest pierwsze objaśnienie (po 10-tym użyciu), cytuję: „MAPE jest uśrednioną wartością APE", gdzie podano formułę dla APE (równanie 15). Następnie na stronie 89 mamy ponownie formułę na APE (równanie 32) oraz formułę dla MAPE (równanie 33), następnie na stronie 94 mamy ponownie zdefiniowane APE (równanie 37-1) i MAPE (równanie 38). Na stronie 203 ponownie przytaczana jest ta sama formuła (równanie 58) z objaśnieniem „MAPE (Mean Absolute Percentage Error) – w przypadku inflacji będzie liczone w następujący sposób". Czasami też pojawia się nietypowy termin, np. w tabeli 46 „Średnia MAPE" (średnia średnich?).

R

Inne miary i zależności są także wielokrotnie definiowane. Praca jest całością, w tekście możemy i powinniśmy wykorzystywać definicje wcześniejsze. Proszę przy wprowadzaniu korekt zadbać o spójność i brak powtórzeń.

Wszystkie przykłady użycia metod są zbyt drobiazgowe. Zbędne i zaciemniające obraz jest pokazywanie kolejnych, drobnych elementarnych przekształceń, przy objaśnianiu metody. Przykładem jest macierz tab. 14 (cała strona danych), która jest pokazywana ponownie w tab. 15 (te same dane po normalizacji), a następnie znowu w tab. 16 (różnice od dobowej wartości przebiegu kroczącego). Jeżeli to ma być przykład obrazujący działanie, to macierz jest za duża (cała strona tekstu), zaś jeżeli to jest wynik analizy danych, to brak jest komentarzy i wniosków. Takie detale można pominąć lub umieścić w dodatku. Przygotowanie danych liczbowych: pobranie danych z bazy, wycinanie kolumn z tabeli itp. to nie jest opis metody. Ten fragment także kwalifikuje się do przeniesienia do dodatku.

Proszę także zredukować treści oczywiste, przykład (str. 131):

> Mianowicie MAPE i APE, pomimo bycia jednostką wyrażaną w procentach, tutaj jest przedstawione w formie dziesiętnej. W celu otrzymania wartości procentowej, każdą przedstawioną wartość w postaci dziesiętnej należy zastosować iloczyn wartości dziesiętnej i 100%. Przykładowo MAPE w formie dziesiętnej 0,1234 odpowiada 12,34 %.

Zakładamy, że czytelnik wie, co to procenty.

4.5 Poprawa redakcji technicznej

Redakcja pracy wymaga korekty. Proszę starać się tworzyć jednolity tekst. Już pierwsza tabela, str 4–6, słownik pojęć, nie jest spójna. Pojęcie ma wyjaśnienie po polsku albo po angielsku, albo po polsku i po angielsku. Wszystkie terminy w tej tabeli mają swoje polskie nazwy, np. autoregression (AR)- autoregresja.

Niektóre elementy (rysunki, tabele) własnych badań są po angielsku, np. Tab. 49 Rys. 85, 86, 87, inne są po polsku.

Skład pracy jest niestaranny. Ok. 20 stron jest częściowo pustych, źle wyrównanych. Rozdziały nie rozpoczynają się od nowej strony. Czcionka na większości rycin jest nieczytelna. Czcionka na rycinach jest różnej wielkości, nawet dla wyników badań własnych. Podobnie tabele, nie jest stosowany spójny rozmiar, ani styl czcionki.

Potrzebna jest korekta, w pracy występują drobne błędy językowe. Sugeruję użycie odpowiednich narzędzi.

4.6 Pozostałe

Proszę dbać o to, aby nie przekazywać informacji nieprawdziwych lub niesprawdzonych. We wstępie (str. 7) pojawia się zdanie: „Pandemia COVID-19 wywołana przez wirus SARS-CoV-2 przyniosła ludzkości więcej ofiar niż jakikolwiek ze znanych kataklizmów

naturalnych, konfliktów zbrojnych oraz pandemii." Z powodu COVID-19 zmarło około 7 milionów osób, epidemia grypy w latach 1920. pochłonęła 100 milionów, II Wojna Światowa 80 milionów. Straty ekonomiczne II Wojny Światowej są znacznie większe niż pandemii COVID-19.

5 Podsumowanie

Rozprawa wymaga wprowadzenia poprawek i ponownego recenzowania. Proponowane poprawki zamieściłem w punkcie 4, dotyczą one formy dostarczonych wyników, która uniemożliwia jednoznaczną ocenę pozytywną. Temat badawczy uważam za bardzo ciekawy i istotny, zaś zaproponowane metody wydają się poprawne, jeżeli tylko Kandydat je właściwie opisze.

Deklaruję możliwość sporządzenia kolejnej recenzji rozprawy mgr. inż. Łukasza Kupracza po wprowadzeniu poprawek i uzupełnień.

KIEROWNIK
Zakładu Sztucznej Inteligencji
dr hab. inż. Robert Nowak
profesor uczelni