Prof. Dr hab. inż. Ewaryst Rafajłowicz

Członek korespondent PAN

Katedra Automatyki, Mechatroniki i Systemów Sterowania

Politechnika Wrocławska

Recenzja rozprawy doktorskiej

mgr inż. Łukasza Kupracza

"Metody predykcji w szeregach czasowych oparte na sztucznej inteligencji w wybranych procesach złożonych"

Niniejsza recenzja została napisana na zlecenie Rady Dyscypliny Informatyka Techniczna i Telekomunikacja w ZACHODNIOPOMORSKIM UNIWERSYTECIE TECHNOLOGICZNYM W SZCZECINIE, w związku z toczącym się przewodem doktorskim Pana mgr inż. Łukasza Kupracza. Promotorem rozprawy jest Pan prof. dr hab. inż. Antoni Wiliński.

Rozprawa liczy 256 stron, w tym bibliografię – 154 pozycje, spis rysunków i wykresów oraz spis tabel.

Tematyka rozprawy

Recenzowana rozprawa dotyczy metod predykcji szeregów czasowych, obserwowanych w złożonych systemach. Jak wynika z zawartości rozprawy, przez złożone systemy Doktorant rozumie procesy o bardzo dużym stopniu skomplikowania, takie jak rozprzestrzenianie epidemii COVID 19 lub zachowanie się poziomu inflacji. Stopień złożoności badanych zjawisk ilustruje też fakt, że mgr Ł. Kupracz rozpatruje je równolegle w wielu krajach świata.

Tak rozumiana tematyka rozprawy rodzi wiele powiązanych ze sobą problemów, do których należą, między innymi,

- zagadnienia pozyskiwania równoległych danych z rozproszonych źródeł,
- ich jakości cyfrowej, porównywalności i wiarygodności merytorycznej,
- bieżącej aktualizacji w celu zapewnienia predykcji opracowywanej w rozsądnym czasie,

- bezpiecznego i trwałego składowania oraz kontrolowanego dostępu do danych,
- dobór metod predykcji lub opracowanie metod i algorytmów predykcji dostosowanych do badanego problemu oraz weryfikacja statystyczna ich dokładności,
- wyspecyfikowanie i uwzględnienie czynników zewnętrznych (zmiennych kontekstowych), które mogą wpływać na jakość predykcji (na przykład gęstości szczepień) i uzupełnienie danych o wartości takich czynników.

Tematykę rozprawy uważam za trafną i aktualną. Trafność tego wyboru dodatkowo uzasadnię w dalszej części recenzji.

1. Krótki przegląd zawartości rozprawy

Rozprawa składa się pięciu rozdziałów i bibliografii

Rozdział wstępny liczy 62 strony i oprócz sformułowania celu rozprawy zawiera bardzo obszerny przegląd literatury na temat rozprzestrzeniania się COVID 19, dostępnych danych i przeglądu podstawowych metod prognozowania, z uwzględnieniem wkładu polskich zespołów badawczych. Rozdział ten ma charakter monograficzny. Jego napisanie wymagało dużego nakładu pracy i sam w sobie stanowi istotny wkład zawarty w ocenianej rozprawie. Rozdział ten zawiera także wstęp do drugiego zadania, mianowicie prognozowania inflacji. Zadanie to jest drugim trudnym zadaniem testowym do badania metod predykcji szeregów czasowych. Doktorant uzasadnia wybór trudnych problemów prognozowania rozwoju COVID 19 i inflacji ich aktualnością i dużą rolą, nie tylko naukową, ale także społeczną. Podobnymi względami kierowało się w ostatnich latach wiele zespołów informatyków i statystyków, które podjęły się prognozowania rozwoju COVID-19.

Rozdział 2 zawiera opis algorytmów predykcji, które są badane dalej w rozprawie. Odnosząc się do nich, będę stosował terminologię, którą przyjął Doktorant. Opisane metody to:

1) Metoda multiregresji z wykorzystaniem pseudoinwersji Moore'a -Penrose'a,
2) Metoda oparta na łańcuchach Markowa,
3) Metoda oparta na podobieństwie dynamiki prognozowanego procesu w zestawieniu z analogicznymi wśród n-najbliższych sąsiednich procesach. Metoda wykorzystuje z wyko-średnią kroczącą,
4) Metoda regresji wielomianem z optymalizacją parametrów uzupełniona korekcją błędów predykcji za pomocą nieliniowej regresji.

5) Metoda regresji z wygładzaniem splajnowym oraz optymalizacją parametrów.

Omawiając w skrócie powyższe metody, zwrócę uwagę na te aspekty, które – moim zdaniem – mają cechy oryginalnego wkładu Doktoranta.

Ad 1) Zaproponowany algorytm bazuje na uogólnionej metodzie najmniejszych kwadratów (MNK), ale Doktorant zaproponował modyfikacje kluczowe dla dostosowania jej do zadań predykcji. Mianowicie, zmiennymi objaśniającymi są wartości szeregów czasowych, które mają przebiegi podobne do prognozowanego procesu. Uogólniona MNK stosowana jest wielokrotnie do objaśniających szeregów czasowych w przesuwającym się oknie czasowym. W ten sposób uzyskiwane są kolejne wartości prognozowanego procesu. Dodatkowym walorem proponowanego podejścia jest uwzględnianie dodatkowych zmiennych kontekstowych. W odniesieniu do prognozowania liczby zachorowań na COVID 19, zmiennymi objaśniającymi (wejściami) są liczby zachorowań w ostatnim wybranym okresie w państwach o podobnym przebiegu epidemii, a zmiennymi kontekstowymi są gęstości szczepień przeciw COVID 19 w tych krajach.

Zaproponowane podejście można zinterpretować także w terminach sieci neuronowej z liniowymi warstwami ukrytymi i przeliczaniem wag na każdym etapie predykcji. Ważna zaletą tego podejścia jest to, że błędy predykcji na popełnione na wcześniejszych etapach nie kumulują się.

Ad 2) Proponowana metoda bazuje na łańcuchach Markowa, ale także w tym przypadku Doktorant musiał zaproponować wybór wektora stanu i sposobu przeliczania stanu na predykowaną wartość fizyczną. W przypadku COVID 19 była to liczba zakażeń spodziewana w następnym etapie predykcji. Doktorant zaproponował empirycznie dobraną formułę przeliczania.

Ad 3) Doktorant zaproponował metodę, o której napisał, że jest „oparta na podobieństwie dynamiki zakażeń wśród n-najbliższych sąsiadach …". W moim przekonaniu, zaproponowane podejście ma znacznie szerszy zakres stosowalności niż tylko predykcja zakażeń. Podejście to można interpretować jako znajdowanie n-najbliższych sąsiadów w odpowiednio zdefiniowanej funkcyjnej przestrzeni metrycznej złożonej z trajektorii podobnych systemów dynamicznych. Następnie, tak jak proponuje Doktorant, uśrednia się te trajektorie i na podstawie przebiegu tej średniej dokonuje się predykcji zachowania trajektorii systemu, który nas interesuje. Ze względu na niepewność danych, Doktorant proponuje obliczanie średniej kroczącej dla każdego obiektu dynamicznego z osobna.

Ad 4) Doktorant opisuje także klasyczne podejście zastosowania modelu wielomianowego względem czasu z dwiema modyfikacjami. Mianowicie, w podstawowym modelu wielomianowym czas jest przesunięty o kilka kroków wstecz, a ich liczba i stopień wielomianu są optymalizowane przy przyjęciu błędu średniokwadratowego jako wskaźnika jakości. Modyfikacja druga polega na wyliczaniu reszt z modelu podstawowego i próbie redukcji sumy kwadratów

reszt za pomocą drugiego modelu wielomianowego względem czasu, lecz tym razem indeks czasu nie jest przesunięty.

Ad 5) Jako piąte z badanych podejść do predykcji, mgr Ł. Kupracz opisuje zastosowanie ekstrapolacji funkcji sklejanych 3 stopnia (qubic splines), które dopasowane są do danych już dostępnych. Proponowane modyfikacje polegają na optymalizacji funkcji sklejanej za pomocą doboru liczby kroków wstecz do punktu, z którego pobierane są dane. Drugą optymalizowaną zmienną jest parametr wygładzania funkcji sklejanej.

W bardzo obszernym (73 strony) Rozdziale 3 Doktorant przedstawia wyniki zastosowania i testowania zaproponowanych metod do predykcji rozwoju COVID 19. Rzadko spotyka się tak szerokie i racjonalnie przeprowadzone badania porównawcze. W celu ich systematyzacji Doktorant zaproponował i szczegółowo opisał udaną adaptację podejścia Goal Question Metric, która jest znana w ewaluacji systemów oprogramowania, ale jej zastosowanie i dostosowanie do celów porównywania algorytmów wymagało nowatorskiego podejścia. W początkowej części Rozdziału 3, Doktorant opisał także autorskie oprogramowanie, które stworzył w celu akwizycji on-line danych o przebiegu pandemii oraz ich weryfikacji, składowania i udostępniania aplikacjom.

Zgodnie z zaprojektowaną procedurą, Doktorant zbadał wpływ gęstości szczepień na dokładność prognoz rozwoju COVID 19 w grupach krajów, w których ta gęstość była istotnie różna. Równie wszechstronne badania zostały przeprowadzone w celu zbadania dokładności prognoz w wybranym kraju w zależności od liczby krajów o podobnej dynamice rozwoju COVID 19, które służą do opracowania prognozy. Do dalszych badań Doktorant wybrał n=7 krajów sąsiednich. Po ustaleniu także innych parametrów, Doktorant przystąpił do systematycznego opisu wyników badań proponowanych algorytmów predykcji. Przedstawił obszerne wyniki badań dla metod, które wyżej omówiłem jako algorytmy 1), 2) i 3). Prezentacje wyników są bardzo dokładne, do tego stopnia, że czytelnik jest w stanie prześledzić rezultaty częściowe dla poszczególnych kroków algorytmów, na przykład macierzy przejść w podejściu 2). Po wszechstronnej ocenie, Doktorant stwierdza, że wśród proponowanych i badanych przez Niego metod, algorytmy bazujące na n-najbliższych sąsiadach (czyli 1) i 3)) prowadzą do znacząco lepszych wyników niż algorytm 2), który opiera się na łańcuchach Markowa. Wskazuje też na wyższość algorytmu multiregresyjnego nad algorytmem 3). **Uzyskane przez Doktoranta rezultaty uważam za cenne i warte dalszego rozwijania.**

Z wielu względów trudno je porównywać z algorytmami prognozowania COVID 19 opracowanymi przez wiele dużych informatycznych zespołów badawczych, często przy współpracy z ośrodkami medycznymi. Część tych algorytmów została opublikowana w NATURE COMMUNICATIONS (2021)12:5173 „A pre-registered short-term forecasting study of COVID-19 in Germany and Poland during the second wave", J. Bracher i ponad 50 współautorów z USA

(MIT, Los Alamos …), Wielkiej Brytanii (Imperial College…), …, Niemiec (Heidelberg, Karslruhe…) I Polski (ICM, MIM UW, MOCOS PWr., …).

Interesujące wyniki na temat COVID 19, które wykraczają poza zakres rozprawy znalazły się publikacji: COVID-19 Pandemic Severity Criterion Based on the Number of Deaths and the Uneven Distribution of These A Wilinski, MK Arti, L Kupracz - IEEE Transactions on Computational Social Systems, 2022, której Doktorant jest współautorem.

Rozdział 3 kończy opis wyników oryginalnego, ale także pracochłonnego podejścia do badania czynników wpływających na dokładności algorytmów predykcji. Mianowicie, Doktorant zbadał wpływ położenia geograficznego grup krajów, w których rozwój COVID 19 był brany pod uwagę w proponowanych algorytmach. Badania objęły 16 krajów na 4 kontynentach. Przedstawione wyniki obszernych badań wskazują, że najmniejszą dokładność uzyskuje się krajach Afryki. Zgadzam się z konkluzją Doktoranta, iż jest to raczej wynik słabej jakości danych z tego obszaru niż cecha algorytmów. Równie ciekawe i pracochłonne badania Doktorant przeprowadził analizując dokładność predykcji w zależności od fazy rozwoju epidemii w wybranych grupach krajów. Predykcji dokonywał za pomocą metod 1) i 3). Dla obu metod najtrudniejsza okazała faza wzrostu zakażeń, a łatwiejsze, ze względu na błąd predykcji, były fazy opadania i stabilizacji zakażeń.

W Rozdziale 4 mgr Ł. Kupracz przedstawił podsumowanie badań dokładności predykcji algorytmów 4) i 5), czyli zmodyfikowanych wersji ekstrapolacji przybliżeń wielomianami i funkcjami sklejanymi. Oprócz wersji zmodyfikowanych Autor przedstawił wyniki dla wersji klasycznych tych podejść. Danymi testowymi były obserwacje poziomu inflacji w 17 wybranych krajach. Również w tym przypadku Autor zastosował opisaną wyżej metodologię testowania. Szczegółowo opisane zostały także miary jakości predykcji, które stosowane były w trakcie badań. Doktorant rzetelnie przedstawił wyniki badań wpływu proponowanych ulepszeń algorytmów, wskazując że w przypadku znacznej części krajów poprawiają one dokładność predykcji, ale w krajach o bardzo wysokiej inflacji i dużej niepewności danych, takich jak Turcja, nie można spodziewać się dokładnych prognoz, niezależnie od użytej metody.

Rozprawę kończy Rozdział 5, zawierający podsumowanie rozprawy i kilka interesujących kierunków badań, które wykraczają poza zakres rozprawy. Podsumowanie jest obszerne i zawiera wnikliwe porównanie badanych algorytmów, ze wskazaniem w jakich sytuacjach który z nich daje dokładniejsze prognozy. Wśród propozycji kierunków dalszych badań część dotyczy potencjalnych ulepszeń proponowanych metod i wskazania innych niż badane w rozprawie obszarów zastosowań. Liczba i różnorodność tych obszarów wskazuje na uniwersalny charakter proponowanych algorytmów. Najciekawsza, ale też najtrudniejsza propozycja Doktoranta dotyczy zbadania wpływu rozwoju pandemii na poziom inflacji. Można się tutaj spodziewać trudności wynikających z niejednoznaczności w odróżnienia wpływu pandemii od decyzji

rządów poszczególnych krajów na zmiany procesów inflacyjnych.

Ocena najważniejszych rezultatów rozprawy.

Głównym celem rozprawy było opracowanie nowych lub zmodyfikowanie znanych algorytmów prognozowania przebiegu szeregów czasowych o złożonej strukturze i zbadanie ich dokładności. Cel ten został osiągnięty, a proponowane podejścia zostały zweryfikowane empirycznie w ponadprzeciętnie szerokim zakresie.

Realizacja tego celu wymagała od Autora rozwiązań bardziej szczegółowych.
Zaliczam do nich:

- Opracowanie i zbadanie 5 algorytmów predykcji. Dwa z nich, a mianowicie algorytm 1), nazwanych przez Doktoranta multiregresją z pseudo-inwersją oraz algorytm 3) czyli n-najbliższych sąsiadów mają istotne cechy oryginalności, mimo że opierają się na klasycznej idei n-najbliższych sąsiadów. Otóż, w klasycznych podejściach najbliżsi sąsiedzi poszukiwani są w ciągu uczącym pochodzącym z tej samej „przestrzeni", której dotyczy decyzja (w tym przypadku – przewidywana wartość szeregu czasowego). Natomiast, propozycje Doktoranta dotyczą prognozowania na podstawie równolegle przebiegających procesów o analogicznej naturze, ale o innym położeniu przestrzennym. W testowanych przykładach rozwoju pandemii COVID 19, te równoległe procesy działy się w innych krajach, niekoniecznie bezpośrednio sąsiadujących. Ponadto, Doktorant opisał sposób doboru tak rozumianych „sąsiadów".

  Algorytm bazujący na łańcuchach Markowa okazał się najmniej dokładny w predykcji zachorowań na SARS Cov2 ze względu na fakt, że brał pod uwagę tylko nieodległe dane z danego kraju. Jednakże i on wymagał od Doktoranta inwencji w dostosowaniu do tej klasy zadań predykcji. Ponadto, w przypadku braku wiarygodnych danych z zewnątrz, może on być jednym z wartościowych podejść. Podkreślić należy, iż mimo testowania na danych z pandemii, proponowane algorytmy mają znacznie szersze zastosowania.
- Pozostałe dwa algorytmy, bazujące na wielomianach i na funkcjach sklejanych również zawierają wartościowe modyfikacje zaproponowane przez Doktoranta i szeroko zbadane na danych dotyczących inflacji.
- Zaproponowanie adaptacji podejścia Goal Question Metric do systematycznego testowania algorytmów i badania ich wrażliwości na czynniki zewnętrzne
- Opracowanie algorytmów akwizycji równoległych danych z rozproszonych źródeł/

Uwagi o charakterze dyskusyjnym

Mam kilkanaście uwag szczegółowych o charakterze redakcyjnym, ale nie utrudniają one oceny treści rozprawy. Dlatego ich tu nie wymieniam, natomiast przekażę je Autorowi w postaci „notatek na marginesach".

Tutaj, chciałbym się skupić na pytaniach i uwagach o charakterze dyskusyjnym.

a) We wstępnej części rozprawy jest wzmianka o kryterium informacyjnym Akkaike (AIC) oraz BIC itp. Czy Doktorant podejmował próby zastosowania ich do badanych w rozprawie przypadków ?

b) Obserwując kolejne fale pandemii COVID 19, nietrudno zauważyć, że przemieszczały się one z kraju do kraju z pewnym, często wielodniowym, opóźnieniem. Przykładowo, w Europie fale te często zaczynały się w Wielkiej Brytanii i „wędrowały" przez Francję i Niemcy do Polski. Zaproponowane w rozprawie algorytmy 1) (multiregresji) i 3) n-najbliższych sąsiadów pozwalają na uwzględnienie takiej dodatkowej informacji. Czy zdaniem Doktoranta mogłoby to wpłynąć na dokładność prognoz ?

Podkreślić należy, że powyższe zagadnienia wykraczają poza ramy rozprawy.

KONKLUZJA

Podsumowując całokształt rozprawy doktorskiej mgr inż. Łukasza Kupracza stwierdzam, że rozprawa ta jest wartościowa, wnosi wkład do aktualnego nurtu badań nad algorytmami predykcji szeregów czasowych oraz nad akwizycją i walidacją danych czasowo-przestrzennych. Ponadto, dzięki dużemu wkładowi pracy Doktoranta, rozprawa ta poszerza naszą wiedzę na temat rozprzestrzeniania się pandemii takich jak COVID 19 i jej związków z inflacją.

W związku z tym stwierdzam, że rozprawa ta spełnia wymagania stawiane ustawowo rozprawom doktorskim i wnoszę o dopuszczenie jej do publicznej obrony.

Prof. Dr hab. inż. Ewaryst Rafajłowicz

Wrocław 3 czerwca 2024 roku