dr hab. inż. Piotr A. Kowalski, prof. AGH
Katedra Informatyki Stosowanej i Fizyki Komputerowej,
Wydział Fizyki i Informatyki Stosowanej,
Akademia Górniczo-Hutnicza w Krakowie,
al. Mickiewicza 30, 30-059 Kraków
email: pkowal@agh.edu.pl

Kraków, 10.08.2021

RECENZJA

rozprawy doktorskiej Pana mgr. Piotra Stanisława Maciąga pt. „Metody odkrywania wzorców sekwencyjnych oraz wykrywania anomalii i predykcji z danych przestrzenno-czasowych ze szczególnym uwzględnieniem ewoluujących impulsowych sieci neuronowych" (Methods of sequential patterns discovery, detection of anomalies and prediction from spatio-temporal data with particular use of evolving spiking neural networks)

1. Uwagi ogólne

Prawną podstawą przygotowania recenzji rozprawy doktorskiej Pana mgr. Piotra Stanisława Maciąga jest Umowa z Politechniką Warszawską – Wydziałem Elektroniki i Technik Informacyjnych reprezentowaną przez Dziekana – Pana Profesora dr. hab. inż. Michała Malinowskiego, którą otrzymałem 11 czerwca 2021 r.

Recenzja została przygotowana na podstawie rozprawy doktorskiej. Przedmiotowa praca została zrealizowana pod kierunkiem Pani prof. dr hab. inż. Marzeny Kryszkiewicz, na Wydziale Elektroniki i Technik Informacyjnych Politechniki Warszawskiej.

Recenzowana rozprawa doktorska przedstawiona została w postaci woluminu wydanego przez Politechnikę Warszawską. Dodatkowo do całości został dołączony dodatek *„Authorship Statements"*, pozwalający na ustalenie udziału procentowego oraz zaangażowania Doktoranta w proces powstawiania poszczególnych publikacji. Ów wolumin został napisany w języku angielskim i składa się 5 części. Pierwszą z nich stanowi wstęp, w którym Autor w pierwszej kolejności opisuje zarówno metody odkrywania wzorców sekwencyjnych jak i procedury wykrywania anomalii i predykcji na podstawie danych przestrzenno-czasowych. W dalszej części skupia się na najczęstszych ograniczeniach tych metod. W końcowej części przedstawione zostały główne tezy badawcze recenzowanej rozprawy oraz przedstawiony został zbiór

publikacji składających się na pracę doktorską. Druga część woluminu została podzielona na dwie sekcje. Pierwsza z nich poświęcona została metodom predykcji i wykrywania anomalii z danych szeregów czasowych przy użyciu ewoluujących impulsowych sieci neuronowych. Autor syntetycznie opisał publikacje P1-P3. W kolejnej sekcji poświęconej metodom odkrywania przestrzenno-czasowych wzorców sekwencyjnych, Doktorant opisał swoje publikacje P4-P8. Kolejną składową zeszytu jest rozdział trzeci, który po krótkim wprowadzeniu bibliometrycznym zawiera kopie publikacji na których opiera się rozprawa doktorska a to: P1-P8. Ostatnią część rozprawy stanowi podsumowanie oraz spis pozycji bibliograficznych, które zostały użyte w recenzowanym dziele. Całość pracy obejmuje 147 stron.

2. Ogólna charakterystyka rozprawy

Rozprawa doktorska Pana Piotra S. Maciąga opiera się na ośmiu publikacjach obejmujących artykuły w renomowanych czasopismach z listy JCR (np. *Environmental Modelling & Software, Neural Networks)* jak również prace będące efektem wystąpień konferencyjnych. W grupie tych ostatnich należy podkreślić jakość owych konferencji, wśród których można odnaleźć *International Joint Conference on Neural Networks*, która jest składową bardzo prestiżowego Światowego Kongresu Inteligencji Obliczeniowej WCCI. Powyższe przekłada się również na punktacje poszczególnych publikacji. W sumie w przedstawionym cyklu wartość tz. „punktów ministerialnych" wynosi 615, a po uwzględnieniu oświadczeń współautorów, wkład Doktoranta wynosi ok. 367 pkt. MNiSW (obecnie MEiN). W tym miejscu warto również podkreślić, iż w cyklu ośmiu publikacji trzy z nich stanowią samodzielne dzieła, które zostały zaprezentowane na znanych konferencjach naukowych (np. *FedCSIS, BDAS*).

Jako zostało wspomniane w pierwszej części rzeczonej recenzji, Doktorant podzielił tematycznie rozprawę na dwie części. W pierwszej z nich, Kandydat podejmuje bardzo ciekawą i ważną społecznie tematykę, proponując nowe metody i algorytmy predykcji zanieczyszczenia powietrza oraz nienadzorowanego wykrywania anomalii w szeregach czasowych. W artykule P1, zaproponowano nowatorski model Clustering-based Ensemble (CEeSNN) do przewidywania zanieczyszczenia powietrza w oparciu o ewoluujące impulsowe sieci neuronowe, które były uczone zindywidualizowanymi szeregami czasowymi poddanymi wcześniej klasteryzacji. W

wyniku symulacji dokonano syntezy predyktora zanieczyszczeń pyłem zawieszonym PM10 oraz ozonu na 1, 3 oraz 6 godzin naprzód bazując na danych kilku stacji w Londynie. Jakość predykcji zaproponowanego modelu CEeSNN, a także modelu singleton NeuCube, sieci MLP i modelu ARIMA została oceniona za pomocą kilku miar jakościowych, a w wyniku weryfikacji zostało stwierdzone, że zaproponowany model jest w stanie dać znacznie lepsze wyniki prognostyczne niż pozostałe trzy modele. W artykule *„Online Evolving Spiking Neural Networks for Incremental Air Pollution Prediction"* tj. P2, kontynuowana jest tematyka zastosowania impulsowych sieci neuronowych zastosowanych w zagadnieniu predykcji zanieczyszczenia powietrza, jednak tym razem autorzy podchodzą do prezentowanej tematyki bardziej teoretycznie. Poszerzając teorię ewolucji impulsowych sieci neuronowych, poprzez nową, szybką i skuteczną technikę kodowania wartości wejściowych do wartości rzędu neuronów wejściowych i wag synaps łączących neurony wejściowe i wyjściowe. Sformułowane zostało również ścisłe górne ograniczenie odległości euklidesowej między wektorami wag synaps neuronu wyjściowego i potencjalnego neuronu wyjściowego, co upraszcza wybór progu podobieństwa stosowanego w fazie uczenia. Tym razem weryfikacja numeryczna obejmowała eksperymenty przeprowadzone na danych dotyczących zanieczyszczeń dla lokalizacji Warszawa-Ursynów. Ostatni z tej serii artykuł P3, poświęcony został dwuetapowemu nienadzorowanemu wykrywaniu anomalii w strumieniu danych z użyciem ewoluujących impulsowych sieci neuronowych, które są używane w trybie on-line. Poza bardzo wnikliwą częścią teoretyczną, imponująca w tym artykule jest cześć eksperymentalna, w której Doktorant razem ze współautorami dokonał porównania jakości proponowanego detektora OeSNN-UAD z 14 innymi detektorami anomalii przedstawionymi w literaturze. Eksperymenty przeprowadzono na plikach danych z dwóch repozytoriów testów anomalii: *Numenta Anomaly Benchmark* i *Yahoo Anomaly Dataset*, które obejmują kilkaset plików. Co więcej, do oceny jakości detektorów anomalii wykorzystano pięć wskaźników: miarę F, precyzję, recall, dokładność zrównoważoną oraz współczynnik korelacji Matthewsa. Warto też nadmienić, iż zawartość merytoryczna w/w badań została opublikowana w bardzo prestiżowym czasopiśmie Neural Networks, które cechuje się 200 punktami MNiSW.

Druga część dysertacji oparta została na artykułach P4-P8, które rozważają zadanie efektywnego odkrywania przestrzenno-czasowych wzorców sekwencyjnych. Pierwszy z artykułów zatytułowany *„A Survey on Data Mining Methods for Clustering Complex Spatiotemporal Data"* ma charakter ogólno-przeglądowy. W publikacji tej Autor podejmuje tematykę typów danych przestrzenno-czasowych oraz metod grupowania danych przestrzenno-czasowych, które oferowane są w literaturze. Duża część pracy poświęcona jest algorytmom dla dwóch problemów już zaproponowanych w literaturze, czyli grupowania złożonych obiektów czasoprzestrzennych jako wielokątów lub obszarów geograficznych oraz mierzenie odległości między złożonymi obiektami przestrzennymi. Kolejny - w tym cyklu - artykuł jest ponownie samodzielną publikacją zatytułowaną *„Efcient Discovery of Sequential Patterns from Event-Based Spatio-Temporal Data by Applying Microclustering Approach"*, która została wydrukowana w wydawnictwie Springer w pracy zbiorowej pod wspólnym tytyłem *„Intelligent Methods and Big Data in Industrial Applications. Studies in Big Data"*. W pracy tej Doktorant rozważa ważne zagadnienie a mianowicie, problem odkrycia wszystkich istotnych wzorców sekwencyjnych oznaczających relacje przestrzenne i czasowe między typami zdarzeń. Artykuł zawiera opracowany autorski efektywny algorytm Micro-ST-Miner, mający zastosowanie do odkrywania przestrzenno-czasowych wzorców sekwencyjnych z wykorzystaniem mikro-grupowania instancji zdarzeń. W przedstawionej procedurze zaadaptowane zostało podejście mikroklastrowe i wykorzystane do skutecznego i wydajnego odkrywania wzorców sekwencyjnych i zmniejszenia rozmiaru zbioru danych instancji. Ponadto zaproponowano odpowiednią strukturę indeksowania i przeformułowano pewne znane już pojęcia. Weryfikacja numeryczna zagadnienia jednoznacznie pokazała przydatność zaproponowanego algorytmu dla generowanych zbiorów danych. Wykazano, że czasy odkrywania wzorców zostały znacznie skrócone oraz pokazano, że rozwiązanie to pozwala na wyeliminowanie wzorców nadmiarowych i szumowych ze zbioru danych. Artykuł P6, będący ponownie samodzielnym dziełem Doktoranta, traktuje o wydajnym wykrywaniu najpopularniejszych wzorców sekwencyjnych w danych przestrzenno-czasowych opartych na zdarzeniach. W artykule tym Doktorant wprowadza notacje sekwencji top-K (wzorzec sekwencyjny), oraz proponuje metodę tworzenia zbioru sekwencji top-K i dynamicznej aktualizacji zbioru na podstawie

sekwencji top-K o zadanej długości. Poprawność zaproponowanej metody tym razem została sprawdzona zarówno na danych syntetycznie wygenerowanych jak i rzeczywistych. Godnym zwrócenia uwagi jest fakt wykorzystania w tej ostatniej grupie, przykładów związanych z zanieczyszczeniami powietrza danymi w postaci siatki zawierającej wartości liczbowe zanieczyszczeń dla regionu Wielkiej Brytanii. Dzięki użytej metodzie możliwym było zbadanie zależności między nietypowymi wystąpieniami zanieczyszczeń w badanym obszarze. Przedostatni artykuł recenzowanego doktoratu stanowi praca, której treścią jest nowatorski algorytm STBFM służący do wykrywania wzorców sekwencyjnych ze zbioru danych instancji zdarzeń i typów zdarzeń. Procedura ta opiera się na strategii wszerz (*breadth-first*) generowania tzw. wzorców kandydyjących. Ciekawym przykładem weryfikacji numerycznej jest zastosowanie rzeczonego algorytmu na zbiorze danych dotyczących incydentów przestępczych dla miasta Boston. Artykuł P8 jest zbliżony tematycznie do P6 i P7 choć niewątpliwie ma znacznie szerszy i ciekawszy wydźwięk szczególnie w aspekcie nowości. W publikacji tej autorzy opisują algorytm mający za zadanie odkrywanie znaczących, zamkniętych przestrzenno-czasowych wzorców sekwencyjnych. W rzeczonym algorytmie wykorzystywany jest wskaźnik uczestnictwa jako miara istotności wykrytych wzorców. Eksperymenty przeprowadzone przy użyciu zbioru danych o zdarzeniach przestępczych w Bostonie, wykazały znaczną konkurencyjność wyników w stosunku do ogólnej liczby znaczących wzorców czasoprzestrzennych, wykrytych przez algorytm STBFM opisany w publikacji P7.

3. Ocena rozprawy

a. Uwagi krytyczno-polemiczne:

1. Uważam, że ciekawym byłoby zbadanie predykcji innych polutantów, które niewątpliwie mają bardzo mocny wpływ na nasze zdrowie. Przykładem takich zanieczyszczeń może być PM2.5 oraz NO_2. Co więcej cześć krajów EU zmaga się bardziej z problemem zanieczyszczenia powietrza poprzez frakcje PM2.5 niż PM10.
2. W opracowaniu modelu prognozy zanieczyszczenia, Autor obliczał wyniki predykcji dla 1, 3 oraz 6 godzin wprzód. Z moich własnych obliczeń wynika, że błąd predykcji stanu zanieczyszczenia powietrza w funkcji czasu przypomina wykres logarytmiczny. Co więcej dla pierwszych kilku godzin model zwykłej regresji

liniowej potrafi bardzo dobrze odzwierciedlić przewidywany stan rzeczywisty. Przyglądając się mierze korelacji Pearsona, dla pierwszej godziny R>0.95 a dla 6 godziny nie spada poniżej 0.85. Zatem czy nie można było by pokazać jak działa zaproponowany przez Doktoranta algorytm prognozy dla np. 24 kolejnych godzin ?

3. W pracy wyraźnie brakuje mi poruszenia bardzo ważnego tematu we współczesnych systemach informatycznych a mianowicie analizy skalowalności proponowanych algorytmów oraz złożoności obliczeniowej czasowej i pamięciowej (z wyjątkiem P4 i P5).
4. W artykule P1, w niewątpliwie bardzo rzetelnej weryfikacji numerycznej Kandydat proponuje kilka miar jakości rozwiązania, a to: MAE, RMSE, IA, R^2. Oczywiście każda z nich ma swój „nośnik informacji" dostarczając nam wiedzę o różnych cechach badanego algorytmu. Czy Doktorant próbował zagregować te oraz inne miary w celu stworzenia jednej uniwersalnej miary? Podobna uwaga dotyczy też artykułów P2 i P3.
5. Na jakiej podstawie Autor proponuje dobór parametrów wewnętrznych do procedur porównawczych takich jak np. ARIMA, MLP w P1; RBF, ARX, MLP, EN w P2 itd.
6. Czy wyniki wersyfikacji numerycznej zaprezentowane w tabelach otrzymano w wyniku pojedynczego procesu uczenie-test, czy może użyto innych metod walidacyjnych?
7. Dość ciekawym jest dobór wektora danych wejściowych w algorytmie prognozy zanieczyszczeń. Dlaczego w pracy Autor używa (P1 wzór 6) dwóch składowych wiatru X i Y? Czy rozważano wzięcie np. max wartości lub długości wektora wypadkowego dla składowych X,Y skoro sam algorytm nie ujmuje zależności przestrzennych (tj. lokalizacji wzajemnych dla badanych stacji) ? Jaka jest fizyczna interpretacja składowych nvPM10 oraz vPM10 oraz czy te dwie składowe nie dają „sumarycznie" PM10 ?
8. Czy rozważane było użycie innych głębokich sieci neuronowych do predykcji zanieczyszczenia?

b. Uwagi szczegółowe

W ramach tego punktu, muszę podkreślić bardzo staranne przygotowanie całego woluminu z pracą doktorską, a przede wszystkim artykułów naukowych. Oczywiście można tu wskazać kilka uwag związanych z pojedynczym brakiem wyjaśnień symboli czy dość skąpym opisem pewnych części pracy. Dla przykładu podam, że rozszerzenia wymagałby opis rysunku 4 w publikacji P5, czy też algorytmów w publikacji P8. Doktorant nie ustrzegł się nielicznych mankamentów natury technicznej takich jak błędy interpunkcyjne czy też pomyłki w skrótach itp. jednak w żadnej mierze nie rzutują one na bardzo wysoką ocenę pracy.

c. Ocena ogólna

Doktorant bardzo dobrze rozumie pojęcie szeregów czasowych zarówno w ujęciu ogólnym, jak i w kontekście danych przestrzenno-czasowych. W szczególności potrafi: (i) opisać ich cechy i własności, (ii) syntetyzować algorytmy pozwalające na głęboką analizę ich własności, (iii) przedstawić wyniki ich analizy, oraz (iv) wybrać i omówić możliwe do zastosowania procedury.

Sposób sformułowania problemu badawczego, przedstawiony w pierwszej i drugiej części, świadczą o dojrzałości naukowej Autora. Analizowany w pracy problem prognostyczny jest bardzo dokładnie sprecyzowany. Mimo mojej krytycznej uwagi w niniejszej recenzji, jestem zdania, iż uzasadnienie użycia impulsowych sieci neuronowych wraz z towarzyszącymi procedurami jest w pełni kompletne. Ponadto bardzo imponującym jest fakt, iż w ośmiu składowych woluminu, trzy z nich są samodzielnymi artykułami Kandydata.

Warto również dodać, że – realizując pracę – Pan mgr Piotr Stanisław Maciąg wykazał się solidnym warsztatem informatycznym. Oprócz zaprogramowania głównych algorytmów, potrafił też umiejętnie wykorzystać wbudowane funkcje następujących zestawów narzędziowych oprogramowania Python oraz MATLAB. Takie umiejętności są niezmiernie istotne, gdyż potwierdzają, że jest On niezależnym naukowcem. Ponadto uważam, że wykonane przez Niego eksperymenty, których liczba jest znaczna, świadczą, że Kandydat ma szczególne predyspozycje do pracy badawczej. Potrafi dobrać odpowiednie wskaźniki oceny jakości modeli prognostycznych,

zilustrować wyniki na rysunkach, przedstawić ważne rezultaty w tabelach i wyciągnąć istotne wnioski. Dodany do pracy opis matematyczny powoduje, że recenzowana przeze mnie rozprawa doktorska jest kompletna i w pełni wartościowa.

4. Podsumowanie

Recenzowana praca doktorska jest przykładem oryginalnego rozwiązania ciekawych zagadnień praktycznych. Do ich rozwiązania Autor rozprawy wykorzystał we właściwy sposób zaawansowane narzędzia technik informacyjnych jakimi niewątpliwie są sieci neuronowe i algorytmy działające na danych przestrzenno-czasowych. Świadczy to o Jego dużej kompetencji w praktycznym posługiwaniu się narzędziami współczesnej informatyki. Niewątpliwie cennym jest umieszczenie części kodów proponowanych algorytmów w ogólnodostępnych repozytoriach, co daje możliwość na wykorzystanie ich zarówno w innych badanych zagadnieniach jak i pozwala na porównanie wydajności i efektywności innych procedur. Uzyskane w pracy wyniki uważam za niewątpliwie oryginalną (nowatorską) propozycję rozwiązania zagadnień technicznych, które zostały wielokrotnie zweryfikowane na drodze rzetelnie przeprowadzanych analiz walidacyjnych. Można zatem uznać, że recenzowana rozprawa doktorska ma charakter oryginalnej pracy projektowo-naukowej, o której mówi bieżąca Ustawa o stopniach naukowych i tytule naukowym oraz stopniach i tytule w zakresie sztuki.

Konkludując uważam, że rozprawa doktorska mgr. Piotra Stanisława Maciąga zdecydowanie spełnia wymagania stawiane w odpowiednich przepisach rozprawom doktorskim i wobec tego stawiam wniosek o jej dopuszczenie do dalszych, przewidzianych Ustawą, etapów przewodu doktorskiego.

Ponadto, biorąc pod uwagę aktualność tematyki badawczej, jej znaczny zakres, wysoką jakość prezentowanych wyników oraz ich istotny wkład w istniejący stan wiedzy i znaczącą aktywność naukową Kandydata, wnioskuję o wyróżnienie recenzowanej rozprawy doktorskiej.

Piotr Kowalski

dr hab. inż. Piotr A. Kowalski, prof. AGH