RADA NAUKOWA DYSCYPLINY
INFORMATYKA TECHNICZNA I TELEKOMUNIKACJA POLITECHNIKI WARSZAWSKIEJ

zaprasz na

PUBLICZNĄ OBRONĘ ROZPRAWY DOKTORSKIEJ
mgr. inż. Piotra MACIĄGA

która odbędzie się w dniu 7 grudnia 2021 roku o godzinie 9:30 w trybie zdalnym.

Temat rozprawy doktorskiej:

„Metody odkrywania wzorców sekwencyjnych oraz wykrywania anomalii i predykcji z danych przestrzenno-czasowych ze szczególnym uwzględnieniem ewoluujących impulsowych sieci neuronowych”

Promotor: prof. dr hab. inż. Marzena Kryszkiewicz - Politechnika Warszawska

Recenzenci: dr hab. inż. Piotr Kowalski - Akademia Górniczo-Hutnicza w Krakowie
prof. dr hab. inż. Tadeusz Morzy - Politechnika Poznańska

* Obrona odbędzie się zdalnie na platformie MS Teams. Osoby zainteresowane uczestnictwem w obronie proszone są o zgłoszenie chęci uczestnictwa w formie elektronicznej na adres sekretarza komisji: Tomasz Gambin– email : tomasz.gambin@pw.edu.pl w dniu 06.12.2021 do godz. 20:00.

Z rozprawą doktorską i recenzjami można zapoznać się w Czytelni Biblioteki Głównej Politechniki Warszawskiej, Warszawa, Plac Politechniki 1.

Streszczenie rozprawy doktorskiej i recenzje są zamieszczone na stronie internetowej: https://bip.pw.edu.pl/Postepowania-w-sprawie-nadania-stopnia-naukowego/Doktoraty/Wszczete-do-30-kwietnia-2019-r/Dyscyplina-informatyka-techniczna-i-telekomunikacja-dziedzina-nauk-inzynieryjno-technicznych/mgr-inz.-Piotr-Maciag.

Przewodniczący Rady Naukowej Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Warszawskiej

dr hab. inż. Jarosław Arabas

Tytuł rozprawy w j. polskim:

Metody odkrywania wzorców sekwencyjnych oraz wykrywania anomalii i predykcji z danych przestrzenno-czasowych ze szczególnym uwzględnieniem ewoluujących impulsowych sieci neuronowych

Streszczenie:

W niniejszej rozprawie zostały przedstawione nowe metody predykcji oraz detekcji anomalii w strumieniach czasowych z wykorzystaniem ewoluujących impulsowych sieci neuronowych oraz nowe metody odkrywania przestrzenno-czasowych wzorców sekwencyjnych. Na rozprawę składa się zbiór ośmiu publikacji, które zostały poprzedzone omówieniem ograniczeń dotychczas zaproponowanych w literaturze metod klasyfikacji i predykcji z wykorzystaniem ewoluujących impulsowych sieci neuronowych oraz odkrywania przestrzenno-czasowych wzorców sekwencyjnych, a także przedstawieniem tezy badawczej.

Pierwsza z publikacji zawartych w rozprawie prezentuje nowy model predykcji zanieczyszczenia powietrza z szeregów czasowych z wykorzystaniem zespołu ewoluujących impulsowych sieci neuronowych. W zaproponowanym rozwiązaniu zespół ten jest trenowany w oparciu o grupowanie danych. Publikacje druga i trzecia zawierają opracowane przez autora rozprawy nowe metody i algorytmy, które stanowią adaptację ewoluujących impulsowych sieci neuronowych uczonych w trybie online do predykcji zanieczyszczenia powietrza oraz nienadzorowanej detekcji anomalii w strumieniach danych.

Publikacje od czwartej do ósmej odnoszą się do problemu efektywnego odkrywania przestrzenno-czasowych wzorców sekwencyjnych. Publikacja czwarta zawiera omówienia typów danych przestrzenno-czasowych oraz metod grupowania danych przestrzenno-czasowych dostępnych w literaturze. Publikacja piąta zawiera nowy efektywny algorytm odkrywania przestrzenno-czasowych wzorców sekwencyjnych z wykorzystaniem mikrogupowania instancji zdarzeń. Publikacje szósta i siódma przedstawiają opracowane: nowy algorytm odkrywania zadanej liczby najbardziej znaczących przestrzenno-czasowych wzorców sekwencyjnych oraz nowy algorytm odkrywania wzorców tego typu wykorzystujący strategię wszerz generowania wzorców kandydujących. W ramach ósmej publikacji został przedstawiony opracowany przez autora rozprawy efektywny algorytm odkrywania wszystkich znaczących zamkniętych przestrzenno-czasowych wzorców sekwencyjnych, stanowiących bezstratną reprezentację wszystkich znaczących przestrzenno-czasowych wzorców sekwencyjnych.

Prof. dr hab. inż. Tadeusz Morzy
Instytut Informatyki
Politechniki Poznańskiej
60-965 Poznań, Piotrowo 2

22.10.2021 r.

RECENZJA ROZPRAWY DOKTORSKIEJ

Methods of Sequential Patterns Discovery, Detection of Anomalies and Prediction from Spatio-temporal Data with Particular Use of Evolving Spiking Neural Networks

Autor rozprawy: Piotr Stanisław Maciąg

1. Jakie zagadnienie naukowe jest rozpatrzone w pracy /teza rozprawy/ i czy zostało ono dostatecznie jasno sformułowane przez autora? Jaki charakter ma rozprawa (teoretyczny, doświadczalny, inny)?

Przedmiotem niniejszej recenzji jest przedłożona rozprawa doktorska zatytułowana „**Methods of Sequential Patterns Discovery, Detection of Anomalies and Prediction from Spatio-temporal Data with Particular Use of Evolving Spiking Neural Networks**" (Metody odkrywania wzorców sekwencyjnych oraz wykrywania anomalii i predykcji z danych przestrzenno-czasowych ze szczególnym uwzględnieniem ewoluujących impulsowych sieci neuronowych), na którą składa się zbiór ośmiu powiązanych tematycznie następujących publikacji:

1. P.S. Maciąg, N. Kasabov, M. Kryszkiewicz, R. Bembenik, Air pollution prediction with clustering-based ensemble of evolving spiking neural networks and a case study for London area, Environmental Modelling and Software, vol. 118, str. 262-280, 2019 (**140 pkt**.)
2. P.S. Maciąg, M. Kryszkiewicz, R. Bembenik, Online Evolving Spiking Neural Networks for Incremental Air Pollution Prediction, Proc. Int. Joint Conference on Neural Networks (IJCNN 2020), str. 1 - 6, 2020, (**140 pkt**.)
3. P.S. Maciąg, M. Kryszkiewicz, R. Bembenik, J.L. Lobo, J. Del Ser, Unsupervised Anomaly Detection in Stream Data with Evolving Spiking Neural Networks, Neural Networks, vol. 139, str. 118-139, 2021 (**200 pkt**.)

4. P.S. Maciąg, A Survey of Data Mining Methods for Clustering Complex Spatiotemporal Data, Proc. of 13th Int. Conference Beyond Databases, Architectures and Structures (BDAS 2017), vol. 716, str. 115-126, 2017 (**20 pkt**.)
5. P.S. Maciąg, Efficient Discovery of Sequential Patterns from Event-Based Spatio-Temporal Data by Applying Microclustering Approach, Intelligent Methods and Big Data in Industrial Applications, vol. 40, str. 183-199, 2019 (**20 pkt**.)
6. P.S. Maciąg, Efficient Discovery of Top-K Sequential Patterns in Event-Based Spatio-Temporal Data, Proc. of Federated Conference on Computer Science and Information Systems (FedCSIS 2018), vol. 15, str. 47-56, 2018 (**20 pkt**.)
7. P.S. Maciąg, R. Bembenik, A Novel Breadth-first Strategy Algorithm for Discovering Sequential Patterns from Spatio-temporal Data, Proc. of the 8th Int. Conference on Pattern Recognition Applications and Methods (ICPRAM 2019), str. 459-466, 2019 (**5 pkt**.)
8. P.S. Maciąg, M. Kryszkiewicz, R. Bembenik, Discovery of closed spatio-temporal sequential patterns from event data, Proc. of the 23rd Int. Conference Knowledge-Based and Intelligent Information & Engineering Systems (KES 2019), str. 707-716, 2019 (**70 pkt**.)

Tematyka badawcza rozprawy doktorskiej Piotra Maciąga, składającej się z powyżej przedstawionego cyklu prac, dotyczy, najogólniej mówiąc, problematyki eksploracji danych, a ściślej, dwóch niezależnych zagadnień w obszarze problematyki eksploracji danych: problematyki odkrywania wzorców sekwencji (sekwencyjnych) w danych przestrzenno-czasowych oraz predykcji i detekcji anomalii w zbiorach szeregów czasowych. Problematyka ta, mimo, że rozwijana już od wielu lat, nadal jest bardzo aktualna i ciągle stanowi obszar aktywnych badań naukowych.

Jak wspomniano już powyżej, opiniowana rozprawa doktorska Pana Piotra Maciąga składa się z 8 prac. W przypadku trzech z nich doktorant jest jedynym autorem, w pozostałych 5 przypadkach doktorant jest jednym z współautorów, przy czym wkład doktoranta wynosi od 45% - 85%. Dwie prace wchodzące w skład rozprawy zostały opublikowane w uznanych czasopismach posiadających IF [P1, P3] (Neural Networks, Environmental Modelling and Software). Jedna praca ukazała się w monografii wieloautorskiej [P5], pozostałe prace [P2, P4, P6, P7, P8] ukazały się w materiałach konferencyjnych (konferencje: IJCNN, KES, BDAS, ICPRAM, FedCSIS). Na szczególne

podkreślenie zasługuje praca [P3], która ukazała się w prestiżowym czasopiśmie Neural Networks (200 pkt).

Pierwsza praca [P1] przedstawia nowy model predykcji zanieczyszczenia powietrza (Clustering-based Ensemble model – CEeSNN) wykorzystujący zespół ewoluujących impulsowych sieci neuronowych (eSNN). W zaproponowanym rozwiązaniu, oryginalne szeregi czasowe (ang. time series) są grupowane w oparciu o wartości opisujące zanieczyszczenie powietrza. Otrzymane w wyniku grupowania klastry szeregów czasowych są, następnie, wykorzystywane do konstrukcji ewoluujących impulsowych sieci neuronowych, tj. pojedynczy klaster szeregów czasowych jest wykorzystany do konstrukcji pojedynczej sieci eSNN. Podstawową kontrybucją przedstawionego podejścia, wg. Autorów, jest wykorzystanie grupowania szeregów czasowych do konstrukcji zbioru treningowego dla zespołu eSNN. Przedstawione w pracy wyniki eksperymentu pokazują, że przedstawione podejście „Cluster-based Ensemble of eSNN" pozwala znacząco poprawić jakość predykcji zanieczyszczenia, mierzonej szeregiem miar jakości, w stosunku do trzech innych podejść stosowanych do predykcji zanieczyszczenia powietrza.

Prace [P2] i [P3] przedstawiają nowy model, nazwany Online evolving Spiking Neural Network for Incremental Prediction (OeSNN-IP), do predykcji i wykrywania anomalii w strumieniach danych. W pracy [P2] przedstawiono zastosowanie zaproponowanego modelu do predykcji zanieczyszczenia powietrza w oparciu o wcześniejsze wartości zanieczyszczenia powietrza zawarte w strumieniu danych opisujących zanieczyszczenie oraz o dane w strumieniu danych pogodowych. W pracy [P3] przedstawiono wykorzystanie zaproponowanego modelu OeSNN do nienadzorowanej detekcji anomalii w jednowymiarowych strumieniach danych. Jest to, wg. Autorów, pierwszy detektor wykorzystujący ewoluujące impulsowe sieci neuronowe uczone w trybie online. Dodatkową kontrybucją pracy jest oryginalna i efektywna technika kodowania danych wejściowych zapewniająca lepszą jakość predykcji aniżeli popularna technika kodowania GRFs stosowana w sieciach eSNN.

Prace [P4-P8] są poświęcone zagadnieniu odkrywania przestrzenno-czasowych wzorców sekwencyjnych. Praca [P4] przedstawia omówienie typów danych przestrzenno-czasowych oraz przegląd metod grupowania danych przestrzenno-czasowych. W pracy [P5] przedstawiono nowy algorytm odkrywania wzorców sekwencyjnych ze zdarzeniowych danych przestrzenno-czasowych (ang. event-based spatiotemporal data) będący modyfikacją algorytmu ST-Miner. Zbiór zdarzeniowych danych przestrzenno-czasowych zawiera zbiór instancji zdarzeń predefiniowanego typu – z każdą instancją zdarzenia jest związany

określony typ zdarzenia, miejsce i czas zajścia zdarzenia. Zaproponowany algorytm wykorzystuje ideę mikrogrupowania, której celem jest redukcja rozmiaru eksplorowanego zbioru danych. Wszystkie instancje zdarzeń tego samego typu, należące do tego samego sąsiedztwa, tworzą mikroklastry, które tworzą indeks mikroklastrów dla eksplorowanego zbioru danych. Idea mikrogrupowania znacząco redukuje czas znajdowania wzorców. Praca [P6] przedstawia analizę i algorytm odkrywania K najbardziej znaczących wzorców sekwencyjnych w zbiorze zdarzeniowych danych przestrzenno-czasowych. W pracy Autor definiuje pojęcie top-K wzorców dla zdarzeniowych danych przestrzenno-czasowych, a następnie, przedstawia algorytm odkrywania takich wzorców sekwencyjnych. Efektywność zaproponowanego algorytmu została zweryfikowana na zbiorach danych syntetycznych i rzeczywistych. W pracy [P7] przedstawiono nowy algorytm odkrywania top-N znaczących wzorców sekwencyjnych wykorzystujący strategię przeszukiwania przestrzeni rozwiązań wszerz. Cenną kontrybucją artykułu jest zaproponowanie struktury Sequential Pattern Tree (SPTree), której celem jest poprawa efektywności procedury generowania zbiorów kandydujących. Efektywność zaproponowanego algorytmu została porównana z efektywnością oryginalnego algorytmu STMiner. Wreszcie, praca [P8] przedstawia nowy algorytm odkrywania zamkniętych przestrzenno-czasowych wzorców sekwencyjnych i stanowi uzupełnienie pracy [P7]. W pracy przedstawiono definicję pojęcia zamkniętego przestrzenno-czasowego wzorca sekwencyjnego oraz przeprowadzono analizę jego własności. Wykazano, że zamknięte przestrzenno-czasowe wzorce sekwencyjne stanowią bezstratną reprezentację przestrzenno-czasowych wzorców sekwencyjnych, oraz przedstawiono algorytm CST-SPMiner odkrywania wszystkich zamkniętych przestrzenno-czasowych wzorców sekwencyjnych.

Jak widać z powyższego, ogólnego, przedstawienia wyników prac składających się na rozprawę, ogólnym celem prowadzonych badań było opracowanie nowych algorytmów predykcji i wykrywania anomalii w szeregach czasowych, z wykorzystaniem zespołu ewoluujących impulsowych sieci neuronowych, oraz opracowanie nowych algorytmów odkrywania przestrzenno-czasowych wzorców sekwencyjnych. Na te ogólne cele rozprawy, jak stwierdza autor w rozdziale 1.3, składają się następujące cele szczegółowe: (1) wykazanie, że można poprawić jakość predykcji zanieczyszczenia powietrza wykorzystując w tym celu zespół ewoluujących impulsowych sieci neuronowych, do konstrukcji których można wykorzystać klastry szeregów czasowych ze zbioru treningowego, (2) wykazanie, że można wykorzystać ewoluujące impulsowe sieci neuronowe uczone w trybie online do efektywnej predykcji danych (w szczególności predykcji zanieczyszczenia powietrza) oraz efektywnej

detekcji anomalii w strumieniach danych, (3) opracowanie efektywnego algorytmu odkrywania przestrzenno-czasowych wzorców sekwencyjnych z wykorzystaniem metody mikrogrupowania oryginalnych danych, (4) opracowanie algorytmu odkrywania top-k najbardziej znaczących przestrzenno-czasowych wzorców sekwencyjnych, oraz (5) opracowanie algorytmu odkrywania zamkniętych przestrzenno-czasowych wzorców sekwencyjnych.

Cele rozprawy i zagadnienie naukowe rozważane w rozprawie zostały jasno sformułowane. Rozprawa ma głównie charakter teoretyczny – opracowanie nowych algorytmów predykcji i wykrywania anomalii w szeregach czasowych, z wykorzystaniem zespołu ewoluujących impulsowych sieci neuronowych, oraz opracowanie nowych algorytmów odkrywania przestrzenno-czasowych wzorców sekwencyjnych. Jednakże, ze względu na intensywny rozwój i popularność narzędzi eksploracji danych, z jednej strony, z drugiej, ze względu na praktyczny aspekt dotyczący np. predykcji danych w szeregach czasowych, w tym predykcji zanieczyszczenia powietrza, wyniki rozprawy mogą być bezpośrednio wykorzystane w praktyce.

2. **Czy w rozprawie przeprowadzono w sposób właściwy analizę źródeł /w tym literatury światowej, stanu wiedzy i zastosowań w przemyśle/ świadczący o dostatecznej wiedzy autora? Czy wnioski z przeglądu źródeł sformułowano w sposób jasny i przekonywujący?**

Uważam, że Autor w sposób właściwy przedstawił stan wiedzy w zakresie algorytmów predykcji i wykrywania anomalii w szeregach czasowych oraz algorytmów odkrywania przestrzenno-czasowych wzorców sekwencyjnych, i posiada dostateczną wiedzę z tego zakresu. Omówieniu aktualnego stanu wiedzy i badań z zakresu problematyki rozprawy są poświęcone rozdziały „Related work" zamieszczonych artykułów, oraz, praca [P4] prezentująca szczegółowy przegląd metod grupowania danych przestrzenno-czasowych. Omówienie stanu wiedzy, przedstawione w rozprawie, jak i bibliografia załączona do artykułów wchodzących w skład rozprawy, świadczą, w mojej ocenie, o **dużej wiedzy autora w zakresie problematyki, której dotyczy rozprawa.**

3. **Czy autor rozwiązał postawione zagadnienie, czy użył właściwej do tego metody i czy przyjęte założenia są uzasadnione?**

Odpowiedź na powyżej postawione pytanie brzmi - **TAK**. Autor rozwiązał poprawnie bardzo trudne problemy w zakresie eksploracji złożonych typów danych, takich jak: szeregi czasowe,

strumienie danych i dane przestrzenno-czasowe. Nie mam zastrzeżeń ani co do przyjętych założeń w rozprawie, ani też co do użytych metod. Przyjęte w rozprawie założenia są typowe dla tego typu zagadnień. Ocena jakościowa zaproponowanych rozwiązań jest również typowa i bazuje na eksperymencie obliczeniowym. Do przeprowadzenia eksperymentu obliczeniowego Autor wykorzystał ogólnie znane i dostępne zbiory testowe.

4. **Na czym polega oryginalność rozprawy, co stanowi samodzielny i oryginalny dorobek autora, jaka jest pozycja rozprawy w stosunku do stanu wiedzy czy poziomu techniki reprezentowanych przez literaturę światową?**

Zbiór prac składających się na rozprawę doktorską Pana Piotra Maciąga jest tematycznie bardzo spójny i dotyczy, jak już wspomniałem, dwóch niezależnych zagadnień w obszarze problematyki eksploracji danych: problematyki odkrywania wzorców sekwencyjnych w danych przestrzenno-czasowych oraz predykcji i detekcji anomalii w zbiorach szeregów czasowych. Uzyskane przez doktoranta wyniki w rozprawie uważam za wartościowe i ciekawe poznawczo. Za podstawową kontrybucję doktoranta uznałbym: (1) wykazanie, że można poprawić jakość predykcji zanieczyszczenia powietrza wykorzystując w tym celu zespół ewoluujących impulsowych sieci neuronowych, (2) wykazanie, że można wykorzystać ewoluujące impulsowe sieci neuronowe uczone w trybie online do efektywnej predykcji danych (w szczególności predykcji zanieczyszczenia powietrza) oraz efektywnej detekcji anomalii w strumieniach danych, (3) opracowanie efektywnego algorytmu odkrywania przestrzenno-czasowych wzorców sekwencyjnych z wykorzystaniem metody mikrogrupowania oryginalnych danych, oraz (4) opracowanie algorytmu odkrywania zamkniętych przestrzenno-czasowych wzorców sekwencyjnych. Do oryginalnej kontrybucji doktoranta zaliczyłbym również ideę wykorzystania grupowania szeregów czasowych do konstrukcji zbioru treningowego dla zespołu eSNN [P1], zaproponowanie oryginalnej i efektywnej techniki kodowania danych wejściowych w sieciach eSNN [P3], oraz ideę wykorzystania mikrogrupowania zdarzeniowych danych przestrzenno-czasowych, która znacząco redukuje czas znajdowania wzorców sekwencyjnych [P5].

Reasumując, uważam, że cele rozprawy, zdefiniowane w punkcie 1.3 zostały w pełni zrealizowane.

5. **Czy autor wykazał umiejętność poprawnego i przekonywującego przedstawienia uzyskanych przez siebie wyników /zwięzłość, jasność, poprawność redakcyjna rozprawy/?**

Autor wykazał umiejętność poprawnego i przekonywującego przedstawienia uzyskanych przez siebie wyników. Przedstawione artykuły są napisane czytelnie, zwięźle i jasno. Poziom edytorski artykułów jest również bardzo dobry. To co jest warte podkreślenia, to duża dojrzałość, jak na młodego naukowca, prezentowanych prac zarówno w odniesieniu do sformułowania rozważanych problemów jak i proponowanych rozwiązań. Jak sądzę, w tym aspekcie, przejawia się istotny wkład współautorów.

6. Jakie są słabe strony rozprawy i jej główne wady?

Jeszcze raz powtórzę, co stwierdziłem już wcześniej, że rozprawa Pana Piotra Maciąga jest ciekawa z naukowego punktu widzenia i wnosi niewątpliwie oryginalną kontrybucję w zakresie problematyki eksploracji danych. Generalnie, nie mam praktycznie żadnych zastrzeżeń do recenzowanej rozprawy. Mam jedną uwagę o charakterze terminologicznym oraz jedną uwagę o charakterze dyskusyjnym w odniesieniu do rozprawy, oraz dwa pytania o charakterze ogólnym dotyczące rozważanej problematyki.

1. **Uwaga dotycząca terminologii polskiej**. Publikacje wchodzące w skład rozprawy zostały przygotowane w języku angielskim, dzięki czemu, generalnie, autor uniknął problemów terminologicznych. Mam jedno zastrzeżenie, które dotyczy krótkiego, polskiego streszczenia znajdującego się na początku rozprawy. W pierwszym zdaniu streszczenia pojawia się termin „strumienie czasowe", który, następnie, w drugim akapicie, zostaje zmieniony na „szeregi czasowe". Oba te terminy odnoszą się do angielskiego terminu „time series". Termin „strumienie czasowe", w moim przekonaniu, jest dosyć niefortunny, gdyż hasło „strumień" jest w jakimś stopniu już zastrzeżone dla angielskiego terminu „data stream". Każdy strumień jest, oczywiście, zmienny w czasie, dodanie przymiotnika „czasowy" niewiele wnosi. Moje zastrzeżenie dotyczy tutaj spójności terminologicznej.
2. **Algorytm odkrywania Top-K wzorców sekwencyjnych**. Moja uwaga w odniesieniu do algorytmu ma charakter wyłącznie dyskusyjny. Nie ujmując nic z zasługi autora w zakresie opracowania algorytmu odkrywania **Top-K wzorców sekwencyjnych** w zbiorze danych przestrzenno-czasowych, chciałbym się odnieść nieco ogólniej do motywacji, która jest przedstawiona w pracy [P6], w kontekście odkrywania Top-K wzorców. Autor stwierdza, że dla wielu zbiorów danych i wielu aplikacji problemem może być racjonalne (sensowne) zdefiniowanie wartości progowej – minimal index sequence threshold dla odkrywanych wzorców sekwencji. To jest oczywiście prawda. Generalnie, w obszarze eksploracji danych, mamy problem definiowania tzw. user-defined parameters (wsparcie, ufność, itp.).

Zaproponowany algorytm odkrywania **Top-K wzorców sekwencyjnych** w zbiorze danych przestrzenno-czasowych stanowi, zdaniem autora, próbę rozwiązania tej wady oryginalnego algorytmu STMiner (uwaga: sekcja V.C - niepoprawna referencja do publikacji, w której zaproponowano algorytm STMiner). Niestety, pojawia się inny problem. Rankingi wzorców, bazujące na miarach „częstości" występowania wzorca (support, density, itp.), borykają się z problemem „dyskryminacyjności" (patrz prace H. Cheng, np. Discriminative Frequent Pattern Analysis for Effective Classification, 2007 IEEE 23rd ICDE), tzn. wzorce o małej częstości są mało dyskryminacyjne, ale również wzorce o dużej częstości (wsparciu) są również mało dyskryminacyjne. W rozważanym tutaj kontekście, dyskryminacyjność wzorca traktuję jako odpowiednik istotności (ang. interestingness) wzorca. W konsekwencji Top-K wzorców może być mało interesujących i mało przydatnych. Być może najciekawsze są wzorce „poniżej" Top-K. Znalezienie Top-K wzorców nie tylko nie rozwiązuje problemu, ale wręcz utrudnia znalezienie interesujących wzorców. Stąd, w ostatnim czasie, w zakresie odkrywania wzorców sekwencyjnych (w zbiorach typu „event sequences") odkrywanie Top-K odnosi się do rankingów opartych o inne miary „ważności" wzorców (ang. interestingness measures) - prace np. L. Feremans et all, N. Tatti (prace znane autorowi, gdyż znajdują się na liście referencji artykułów [P6] i [P7]).

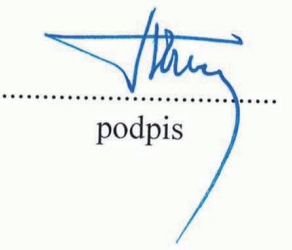
3. **Wzorce sekwencyjne z ograniczeniami w danych przestrzenno-czasowych.** Uogólnione sformułowanie problemu odkrywania wzorców sekwencyjnych w sekwencyjnych bazach danych zakłada możliwość definiowania różnego rodzaju ograniczeń w odniesieniu do odkrywanych wzorców sekwencyjnych. Szczególnym przypadkiem takich ograniczeń są ograniczenia czasowe nakładane na odstępy czasowe pomiędzy wyrazami sekwencji – tzw. gap constraints. Czy sensowne jest definiowanie tego typu ograniczeń w odniesieniu do wzorców sekwencyjnych odkrywanych w zbiorach danych przestrzenno-czasowych?
4. **Wyzwania w obszarze odkrywania wzorców sekwencyjnych w danych przestrzenno-czasowych**. W popularnym przeglądzie „Spatio-Temporal Data Mining: A Survey of Problems and Methods, G. Atluri, A. Karpatne, V. Kumar, ACM Computing Surveys 51, 2018, jako dwa podstawowe wyzwania w obszarze odkrywania wzorców sekwencyjnych w danych przestrzenno-czasowych autorzy wskazują: „Some of the key challenges in mining ST sequential patterns include defining interesting measures that capture meaningful non-spurious patterns and developing efficient approaches to discover interesting patterns from an exponentially large space of candidate patterns".

Recenzowana rozprawa wpisuje się w rozwiązanie drugiego z wymienionych wyzwań. Moje pytanie dotyczy pierwszego z wymienionych wyzwań. Jaki jest zasadniczy „zarzut" odnośnie miary „sequence index", która jest wykorzystywana w rozprawie jako „interesting measure"?

Podsumowanie:

Uważam, że cele rozprawy, zdefiniowane w punkcie 1.3, zostały w pełni zrealizowane. Autor przedstawił w rozprawie nowe algorytmy predykcji i wykrywania anomalii w szeregach czasowych, z wykorzystaniem zespołu ewoluujących impulsowych sieci neuronowych, oraz nowe, oryginalne algorytmy odkrywania przestrzenno-czasowych wzorców sekwencyjnych. Efektywność zaproponowanych rozwiązań wykazał na podstawie szeroko przeprowadzonych eksperymentów obliczeniowych. Uzyskane przez doktoranta wyniki w rozprawie uważam za oryginalne, wartościowe i ciekawe poznawczo.

Stwierdzam zatem, że recenzowana rozprawa doktorska Pana Piotra Maciąga spełnia z nadmiarem wymagania stawiane rozprawom doktorskim przez obowiązującą ustawę i wnoszę o dopuszczenie jej do publicznej obrony. Ze względu na oryginalną kontrybucję rozprawy oraz jakość uzyskanych wyników wnoszę również o wyróżnienie opiniowanej rozprawy.



.....

podpis

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