**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]).

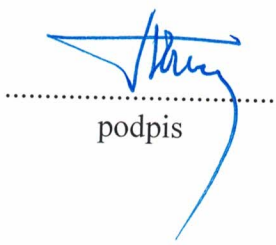
3. **Wzorce sekwencyjne z ograniczeniami w danych przestrzenno-czasowych.** Uogólnione sformułowanie problemu odkrywania wzorców sekwencyjnych w sekwencyjnych bazach danych zakłada możliwość definiowania różnego rodzaju ograniczeń w odniesieniu do odkrywanych wzorców sekwencyjnych. Szczególnym przypadkiem 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