dr hab. inż. Artur Gramacki, prof. UZ
Uniwersytet Zielonogórski
Wydział Informatyki, Elektrotechniki i Automatyki
Instytut Sterowania i Systemów Informatycznych
Zielona Góra, 65-516
ul. prof. Zygmunta Szafrana 2

Zielona Góra, dn. 11 września 2022 r.

Recenzja rozprawy doktorskiej
mgr. inż. Tomasza Rybotyckiego
pt. „Estymacja gęstości rozkładu niestacjonarnych danych strumieniowych"
napisanej pod kierunkiem prof. dr hab. inż. Piotra Kulczyckiego

1. Podstawa opracowania recenzji

Podstawą opracowania recenzji jest pismo z-cy Dyrektora ds. Naukowych Instytutu Badań Systemowych Polskiej Akademii Nauk dr. hab. inż. Jana Owsińskiego z dnia 16 maja 2022 r. w sprawie powołania mnie na recenzenta rozprawy doktorskiej mgr. inż. Tomasza Rybotyckiego.

2. Zakres rozprawy doktorskiej

Tematyką recenzowanej pracy doktorskiej jest analiza danych mających charakter strumieniowy. Autor przedstawia w niej pewien opracowany przez siebie algorytm (nazwany akronimem DEDSTA, od *Density Estimation for Data Stream with possible Trend Algorithm*) oraz bada jego praktyczną przydatność.

3. Ogólna ocena struktury i zawartości rozprawy

Doktorant niestety nie stawia w swojej dysertacji doktorskiej tezy, której poprawność (prawdziwość) będzie starał się wykazać w rezultacie realizacji swojej rozprawy doktorskiej. Takie podejście, co do formy pisania rozprawy doktorskiej, jest klasyczne i powszechnie stosowane. Pewną namiastkę tezy można odnaleźć jedynie w ostatnim akapicie rozdziału pierwszego.

Ponadto powszechnie praktykowanym zwyczajem jest udostępnienie, wraz z głównym tekstem rozprawy doktorskiej, jego skróconej, syntetycznej wersji w postaci autoreferatu. Autor niestety takiego opracowania nie przygotował a udostępnione półstronicowe streszczenie na pewno na miano autoreferatu nie zasługuje.

Recenzowana rozprawa doktorska liczy, łącznie z bibliografią, 76 stron. Jest to zatem opracowanie objętościowo niewielkie – oczywiście o końcowej ocenie przesądza przede wszystkim zawartość merytoryczna. W spisie literatury Autor umieścił 79 pozycji a dokonany wybór jest zasadniczo zgodny z głównym tematem rozprawy. Brakuje tam moim zdaniem ważnych publikacji na temat zaawansowanych metod wyboru tzw. współczynnika wygładzania dla estymatorów jądrowych. Dokładniej piszę o tym w rozdziale 4.

Należy w tym miejscu zauważyć, że dorobek publikacyjny doktoranta, który niejako wieńczy doktorat (a tak przynajmniej zwykle bywa i być powinno), jest mim zdaniem niewystarczający. W spisie literatury zamieszczone są 3 pozycje z afiliacją Autora. Jedna z nich (pozycja [57]) dotyczy zupełnie innych niż omawiane w rozprawie zagadnienia i wg mnie nie powinna być w spisie literatury użyta. Z kolei pozycja [48] moim zdaniem nie powinna znaleźć się w ogóle w spisie literatury, gdyż dotyczy artykułu w trakcie publikowania. Nie ma żadnej informacji w jakim dokładnie czasopiśmie artykuł został zgłoszony, a ponieważ Autor używa tu zwrotu „w trakcie publikowania" (bardziej poprawnie chyba powinno to być

oznaczone jako „wysłane do czasopisma") nie sposób wyrokować w tym momencie o jej dalszych losach. Jedyną więc publikacją *stricte* zgodną z tematyką rozprawy jest artykuł konferencyjny opublikowany w ramach uznanej w świecie naukowym konferencji ICCS 2021 (International Conference on Computational Science, pozycja [47]).

Pracę rozpoczyna rozdział zatytułowany przez Autora „Przedmowa". Po zapoznaniu się z jego zawartością wydaje mi się, że zaprezentowane tam treści powinny znaleźć się raczej w kolejnym rozdziale zatytułowanym „Wstęp". Podając za Wielkim Słownikiem Języka Polskiego: przedmowa to „tekst na początku książki na temat jej treści, charakteru czy okoliczności wydania, nieraz także z podziękowaniami lub dedykacją, sporządzony przez autora, tłumacza lub wydawcę". Zaprezentowana „Przedmowa" zupełnie nie odpowiada powyższemu.

Kolejne rozdziały, a więc „Przegląd literatury" oraz „Preliminaria matematyczne" są swego rodzaju kompilacją dostępnej literatury przedmiotowej. W kolejnym rozdziale zatytułowanym „DEDSTA – predykcyjna estymacja gęstości danych strumieniowych" Autor przedstawia opracowany przez siebie algorytm do realizacji zadania jak w tytule rozdziału. Jest to, jak wydaje się, najważniejszy merytorycznie rozdział w całej rozprawie. W rozdziale pt. „Wybrane metody alternatywne" intencją Autora było pokazanie, iż na przestrzeni wielu lat naukowcy opracowali również inne niż oparte o estymatory jądrowe metody estymowania funkcji gęstości dla danych strumieniowych. Przedostatnim rozdziałem rozprawy jest rozdział zatytułowany „Weryfikacja numeryczna", gdzie Autor stara się pokazać poprawność działania oraz skuteczność swojego algorytmu. Pracę pisemna kończy „Podsumowanie". Różne uwagi szczegółowe dotyczące poszczególnych rozdziałów zostaną sformułowane w dalszej części recenzji.

4. Uwagi recenzenta

Uważna lektura rozprawy doktorskiej pozwala sformułować pewne uwagi natury ogólnej oraz uwagi bardziej szczegółowe. Poniżej ich zestawienie.

Uwagi ogólne

Autor opiera zasadniczo całą swoją rozprawę na jednym algorytmie przedstawionym w rozdziale czwartym. Algorytm ten w zasadzie w całości ma charakter heurystyczny. Pojawia się tam bardzo wiele różnego rodzaju założeń, które jak się wydaje są przyjęte w pewnym sensie na wiarę. Taki sposób postępowania zapowiedziany jest niejako we wstępie pracy (strona 9), gdzie Autor stwierdza, że (cytuję) „Proponowana koncepcja ma charakter kompleksowy i nie wymaga dogłębnych teoretycznych dywagacji ani żmudnych badań". Oczywiście to, że algorytm jest w przeważającej większości heurystyczny samo w sobie nie jest czymś nagannym, znanych jest przecież sporo takich algorytmów w różnych dziedzinach nauki. Poprawność i przede wszystkim praktyczna skuteczność tego typu algorytmów wymaga jednak bardzo szczegółowych badań, głównie eksperymentalnych. Niestety Autor zadanie to wykonał w sposób niezadawalający. Algorytm w zasadzie został przetestowany na jednym, dość specyficznym, sztucznie wygenerowanym zbiorze danych mającym charakter danych strumieniowych. Strumień ten to tak *de facto* dwa liniowo przyrastające strumienie, skok jednostkowy oraz dane o ustalonej amplitudzie (strona 46, wzory 6.1 oraz 6.2). Wszystko to jest zaburzone szumem o standardowym rozkładzie normalnym. Trudno taki zestaw eksperymentalny uznać za oddający wystarczająco dobrze rzeczywiste strumienie danych spotykane w praktyce. Wyciąganie więc wiążących wniosków na temat praktycznej skuteczności algorytmu jest wątpliwe. Co prawda algorytm został uruchomiony również na trzech rzeczywistych danych strumieniowych (temperatura oraz wilgotność w trzech wybranych miastach, patrz strona 54 i kolejne). Dla danych tych oczywiście nie są znane prawdziwe rozkłady gęstości, więc można tylko domniemywać, czy otrzymane wyniki są poprawne (z praktycznego punktu widzenia). Autor analizując otrzymane wyniki podaje dość oczywiste

wnioski, które w zasadzie można wyciągnąć wprost analizując (nawet wzrokowo) odpowiednie przebiegi czasowe. Przykładowe wnioski to (cytuję) „Analiza wyników ... wskazuje, że w marcu 2017 roku średnia temperatura w Minneapolis oscylowała w otoczeniu 0 stopni", czy też „Na początku lata średnie temperatury były już wyraźnie wyższe i stabilnie utrzymywały się na poziomie około 11°C". Trudno uznać te wnioski za odkrywcze.

Oddzielnego omówienia wymaga kwestia zastosowania omawianego algorytmu i otrzymane wyniki dla danych strumieniowych wielowymiarowych. Autor w kilku miejscach stwierdza, że opracowany algorytm ma charakter uniwersalny i jego stosowalność nie ogranicza się jedynie do danych jednowymiarowych. Niestety zamieszczone w pracy wyniki eksperymentalne dla tego rodzaju danych są jedynie śladowe (strony 66-68) a wyciągnięte wnioski praktycznie bez żadnego znaczenia. Zamieszczone na kilku rysunkach wykresy gęstości prawdopodobieństwa (w postaci wykresów poziomicowych) kompletnie nic nie wnoszą. Poziomice są tak „poszarpane", że w zasadzie nie jest możliwe wyciągnięcie na ich podstawie jakichkolwiek sensownych wniosków. Podejrzewam, że Autor zupełnie niewłaściwie dobrał współczynnik (lub współczynniki, bo nic na ten temat nie ma w tekście pracy) wygładzania dla estymatora jądrowego. Autor niestety nie przeprowadził żadnych eksperymentów ze sztucznie wygenerowanymi danymi 2D, gdzie znana byłaby gęstość prawdopodobieństwa. Aby uwiarygodnić rzeczywistą wielowymiarowość algorytmu należałoby również wykonać eksperymenty na danych więcej niż 2D. Oczywiście w takiej sytuacji należałoby zaproponować inną niż graficzna formę prezentacji wyników (dla danych więcej niż 3D).

Uwagi bardziej szczegółowe

Autor podaje w pracy linki do stron internetowych z używanymi danymi rzeczywistymi (rozkłady temperatur i wilgotności dla Minneapolis, Rio de Janeiro oraz Krakowa). Niestety linki te nie prowadzą do wspomnianych danych. Autor powinien jednak te dane sprawdzić dokładniej lub (tak byłoby najrozsądniej) umieścić na przygotowanej stronie internetowej (pozycja [69] w spisie literatury).

Przygotowana strona internetowa, będąca suplementem do dysertacji doktorskiej, jest bardzo minimalistyczna. Są tam umieszczone wyłącznie linki do wyników eksperymentalnych bez żadnych komentarzy. Wyniki stanowią pliki graficzne (rozszerzenie *png*) prezentujące wyniki w każdym kroku obliczeń. Szkoda, że pliki te nie są przygotowane dodatkowo w skalowalnym formacie wektorowym, gdyż czasami, gdy chodzi o zaobserwowanie rożnych subtelności, nieskalowalna postać pikselowa zawodzi. Niezrozumiałe jest też umieszczanie wszystkich tych plików „jak leci" zamiast tych najciekawszych, czy też najważniejszych i obowiązkowo z jakimiś komentarzami. W obecnej wersji strony mamy tam niemal 100.000 (!) pojedynczych plików graficznych, których nie sposób przecież przejrzeć.

Wydaje się, że należało również pomyśleć o przygotowaniu oprogramowania (może wraz z kodami źródłowymi), które pozwoliłyby czytelnikowi samodzielnie przeprowadzić dodatkowe własne eksperymenty. Ta uwaga jest szczególnie istotna wobec zastrzeżeń podanych wyżej a dotyczących moim zdaniem zbyt małej liczby eksperymentów numerycznych potwierdzających praktyczną przydatność algorytmu – zarówno jeżeli chodzi o dane jednowymiarowe, jaki i szczególnie wielowymiarowe.

Wiadomym jest, że w dziedzinie jądrowych estymatorów funkcji gęstości prawdopodobieństwa (również ich pochodnych) jednym z najistotniejszych elementów jest właściwy wybór parametru wygładzania, bądź to w postaci skalara (zwykle oznaczanego h), bądź też macierzy (zwykle oznaczanej H). Niewłaściwy ich wybór może zrujnować wyniki końcowe, przykładem niech będą wyniki zamieszczone w rozdziale 6 a dotyczące analiz danych strumieniowych 2D. Należało chyba dokładniej przejrzeć aktualną literaturę na ten temat, dużo w tej dziedzinie dokonał np. dr Tarn Duong

(https://www.mvstat.net/tduong/), który bardzo kompleksowo opracował to zagadnienie i wyniki zaprezentował w cyklu kilku artykułów. Zamieszczona na stronie 17 informacja, iż algorytm *plug-in* ma złożoność kwadratową, w świetle wspomnianych prac (oraz pracy [23]), nie jest już w zasadzie aktualna. Wykorzystując w odpowiedni sposób algorytm FFT można obniżyć złożoność obliczeniową do akceptowalnych wartości ($O(N \cdot \log_2 N)$). W przypadku zastosowania algorytmu DEDSTA do bardzo licznych zbiorów danych (a przecież dane strumieniowe niejako z definicji takie są) może mieć to niebagatelne znaczenie.

Nieco dyskusyjne jest dość kategoryczne stwierdzenie (strona 15, użyto zwrotu „najczęściej zaleca się"), iż konstruując estymator jądrowy powinno się głównie wybierać jądro produktowe (czyli *de facto* iloczyn *n* jąder jednowymiarowych, oddzielnie dla każdej *n*-tej współrzędnej). Jądro radialne ma chyba jednak lepsze właściwości i skonstruowany na jego bazie estymator jądrowy lepiej potrafi dopasować się do *n*-wymiarowych danych, zwłaszcza gdy użyty parametr wygładzania ma postać macierzy *H* pełnej (czyli, gdy macierz *H* nie jest diagonalna).

Jak już wyżej wspomniano algorytm DEDSTA bazuje na wielu heurystycznych założeniach. Część z nich jest w akceptowalny sposób uwiarygodniona, jednak dla sporej liczby założeń nie odnajduję zadawalającego uzasadnienia. Poniżej podaję wybrane przykłady.

Strona 19, przyjęcie wartości r = 0.01, 0.05, 0.1 jako (cytuję): „w praktyce używa się najczęściej".

Strona 21, użycie 7-krotności najmniejszego poziomu krytycznego 0.01. Do wyznaczenia tej wartości Autor dość enigmatycznie wskazuje tu na metodę Zieglera-Nicholsa używaną powszechnie do automatycznego doboru nastaw regulatorów PID. Chyba należałoby nieco dokładniej przybliżyć użytą tu argumentację, najlepiej popartą konkretnymi wyliczeniami.

Strona 27, przyjęcie we wzorze 4.4 stałej wartości 1.1. Dalej w tekście pojawia się próba uzasadnienia przyjęcia takiej a nie innej wartości ale jest ona prawdę mówiąc dość zawiła i nie do końca przekonywująca.

Strona 28, przy ustalaniu wartości m_0 stwierdzono (cytuję): „jeżeli gęstość rozkładu badanego strumienia jest funkcją istotnie wielomodalną, to dobrą praktyką może być zwiększenie tej wartości o 100 na każdy dodatkowy mod". Dlaczego akurat 100 a nie na przykład 80 lub 150? Wydaje mi się ponadto, że w eksperymentach numerycznych, których wyniki podano w pracy, nie badano danych, gdzie gęstość rozkładu jest wielomodalna.

Strona 28, wzór 4.5 oraz stwierdzenie (cytuję): „Można zatem traktować wartość m_{min} = 100 jako standardową."

Strona 29, (cytuję): „Należy również dodać, ze nieliniowe postacie formuły (4.6), np. kwadratowa czy też pierwiastkowa, nie dawały korzystniejszych rezultatów, jakby nadmiernie pomniejszając informację reprezentowaną przez dawne (postać kwadratowa) lub nowsze (formuła pierwiastkowa) elementy rezerwuaru.". W tekście pracy nie odnajduję żadnego śladu eksperymentów dotyczących formuły 4.6.

Strona 31, (cytuję): „Do większości zastosowań można zaproponować wartości od 1/3 do 2/3, przy czym dla wolniejszych zmian, preferowana jest mniejsza z tych wartości, natomiast większa dla szybkich".

Na stronach od 31 do 35 pojawiają się sformułowania, przykładowo: „ostatecznie proponowane jest", „należy przyjąć", „powinny być użyte", można zaproponować także", „powinno być przyjęte w postaci", „należy arbitralnie przyjąć h=1", „warto wówczas utrzymać h=1", „wartości parametrów m_0, m oraz m_{min} powinny być pomnożone przez czynnik 4^n".

Jak wspomniano wyżej w tekście pracy jest więcej tego typu heurystyk, aby nadmiernie nie wydłużać recenzji, nie podaję ich wszystkich ograniczając się tylko do tych najważniejszych.

Podkreślę w tym miejscu ponownie, iż nie uważam algorytmów heurystycznych za coś złego. Często złożoność zagadnień jest tak duża, że nie ma faktycznie możliwości wszystko analitycznie wykazać, czy też udowodnić. Niemniej jednak wersyfikacja praktycznej użyteczności takich algorytmów bezwzględnie wymaga wielu eksperymentów z użyciem możliwie jak najróżnorodniejszych danych testowych. Ten element w recenzowanej pracy został niestety zrealizowany w stopniu niewystarczającym.

Inne pomniejsze uwagi

Autor w tekście pracy dość dowolnie używa form w czasie przyszłym, teraźniejszym i przeszłym. Ponieważ mówimy o podsumowaniu pewnych prac już wykonanych i pewnym sensie zakończonych, raczej należałoby używać czasu tylko przeszłego.

Niezbyt szczęśliwie wybrano oznaczenia dla liczności rezerwuaru (tak Autor nazywa szerokość ruchomego okna). Skoro minimalna liczność tegoż rezerwuaru oznaczona została jako m_{min}, to dlaczego wartość maksymalna oznaczona zostało jako m_0? Zero raczej słabo kojarzy się z wartością maksymalną.

Autor niezbyt precyzyjnie (często chyba zamiennie) używa zwrotów: „element nietypowy", „element rzadki", „element odstający", „element oddalony". Zdecydowanie należałoby to uściślić.

W rozdziale 3.4 (strona 20) Autor pisze, że (cytuję) „test KPSS posiada dwie wersje: zakładającą istnienie trendu lub nie". Chodzi chyba raczej o wersje (użyję tutaj oryginalnych angielskich nazw): *level stationary* oraz *trend stationary*?

Na stronie 21 Autor pisząc o użyciu statystyki KPSS do zadania określenia stopnia niestacjonarności powołuje się na (cytuję) „aksjomaty logiki rozmytej". W tekście pracy nie ma „ciągu dalszego" i wyjaśnienia, jak należy to rozumieć i jaki był dokładnie tok rozumowania.

Na stronie 25 Autor stwierdza, że (cytuję): „opracowana procedura umożliwia identyfikację elementów nietypowych, z zaznaczeniem które z nich reprezentują tendencje nowopowstałe, a które wygasające." W tekście pracy nie znalazłem dokładnego wyjaśnienia, może nawet jakiejś formalnej definicji, jaka jest różnica między tymi dwoma rodzajami elementów nietypowych. Na tej samej stronie autor pisze (cutuję): „Pozwala to także określić które z elementów nietypowych są związane ze zjawiskami wstępującymi, a które z regresywnymi". Podobnie jak poprzednio te sformułowania nie są jasno wyjaśnione. Z kolei na stronie 33 Autor (chyba w sposób niezamierzony) używa określeń „elementy nietypowe z tendencją wzrostową" oraz „elementy recesywne".

W eksperymentach numerycznych pojawia się stwierdzenie, iż w jakimś momencie działania (cytuję): „następuje konsolidacja algorytmu". Jak należy to rozumieć? Bez jakiegoś dokładniejszego wyjaśnienia trudno się domyśleć istoty tej konsolidacji.

Na wszystkich rysunkach w rozdziale 6 (ale też i tych udostępnionych na stronie internetowej) pojawiają się numery rozdziałów od 3.1 do 3.4. Chodzi zapewne o rozdziały 4.1 do 4.4. Prawdopodobnie zmieniono w jakimś momencie układ rozdziałów a nie poprawiono rysunków (prawdopodobnie wymagałoby to przegenerowania wszystkich obliczeń, aby utworzyły się nowe rysunki).

Na rysunku 6.1 (ale też i na innych niezamieszczonych bezpośrednio w pracy) pojawia się bardzo duża wartość parametru KPSS (np. ponad 8). W świetle rozdziału 3.4 wydaje się, że ta wartość powinna zawsze należeć do przedziału [0,1]. Czy więc nie ma jakiś błędów w programie komputerowym? Należałoby to wyjaśnić.

Na stronie 49 w uwagach na temat rozwinięcia koncepcji algorytmu DEDSTA na dane wielowymiarowe pojawia się fragment (cytuję): „Można skonstatować, że zmiany wolniejszych atrybutów nie mają istotnego wpływu na jakość estymacji". Co dokładnie oznacza „wolny atrybut"?

W podsumowaniu (strona 70) niektóre sformułowania dotyczące przyszłych badań w przedmiotowej tematyce brzmią bardzo enigmatycznie, żeby nie powiedzieć niezrozumiale, przykładowo (cytuję): „Kolejnym przedmiotem prac będzie ujęcie warunkowe, w ramach którego możliwe jest uwzględnienie pomiarów tych wielkości, których aktualna wartość pozwala na istotne uściślenie stosowanego modelu probabilistycznego."

W ostatnim zdaniu podsumowania Autor wymienia z imienia i nazwiska osoby, które prawdopodobnie w przyszłości zajmą się podobną tematyką i ew. rozwiną różne zagadnienia zawarte w rozprawie. Chyba nie ma w zwyczaju tak czynić.

5. Konkluzja końcowa

Recenzowana praca dotyczy dość istotnego zagadnienia naukowego. Autor zapewne włożył sporo wysiłku w opracowanie będącego podstawą rozprawy doktorskiej algorytmu DEDSTA. Zabrakło chyba jednak czasu (lub inne czynniki na to wpłynęły), aby wszystko dokładnie dopracować i uzupełnić do takiego stopnia, aby dysertacja mogła być uznana za dzieło skończone. Recenzowana rozprawa niestety wydaje się być pracą niedokończoną, aczkolwiek z potencjałem na przyszłość.

Biorąc więc pod uwagę powyższe, jak i wcześniej sformułowane oceny cząstkowe (w sporej części negatywne) stwierdzam, że recenzowana rozprawa doktorka nie spełnia wymogów ustawowych stawianych pracom doktorskim i w związku z tym nie mogę wnioskować o dopuszczenie mgr. inż. Tomasza Rybotyckiego do dalszych etapów przewodu doktorskiego.

Artur Gramacki