POLITECHNIKA WARSZAWSKA
WYDZIAŁ MATEMATYKI i NAUK INFORMACYJNYCH
ul. Koszykowa 75
00-662 WARSZAWA
NIP: 525-000-58-34
Tel.: (22) 234-79-88, 621-93-12
Fax.: (22) 625-74-60

Warszawa, 28 lipca 2022 roku

Prof. dr hab. inż. Jacek Mańdziuk
Wydział Matematyki i Nauk Informacyjnych
Politechnika Warszawska

Recenzja rozprawy doktorskiej mgr inż. Tomasza Rybotyckiego zatytułowanej *Estymacja gęstości rozkładu niestacjonarnych danych strumieniowych*

Recenzja została przygotowana na prośbę Zastępcy Dyrektora Instytutu Badań Systemowych Polskiej Akademii Nauk, dr hab. inż. Jana W. Owsińskiego, wyrażoną w piśmie z dnia 16 maja 2022 roku.

Tematyka rozprawy

Przedstawiona do recenzji rozprawa dotyczy zagadnienia estymacji gęstości rozkładu danych strumieniowych, ze szczególnym uwzględnieniem strumieni niestacjonarnych. W przypadku tego typu danych klasyczne metody estymacji rozkładu często zawodzą z uwagi na brak możliwości składowania tak dużej ilości danych oraz z racji na ich zmienny w czasie charakter. W ramach prowadzonych przez Doktoranta badań powstała autorska metoda estymacji gęstości rozkładu polegająca na adaptacyjnym doborze liczby najświeższych danych strumienia stanowiących jego aktualną reprezentację oraz posiadająca możliwość aktualizacji estymowanej gęstości w oparciu o napływające dane. Kluczowym aspektem proponowanego rozwiązania jest także bieżąca analiza trendów oraz wykrywanie elementów nietypowych.

Skuteczność zaproponowanej metody została zweryfikowana eksperymentalnie w oparciu o dane syntetyczne (jednowymiarowy szereg czasowy) oraz rzeczywiste (dane meteorologiczne z Minneapolis, Rio de Janeiro oraz Krakowa).

Hipoteza badawcza

Nie znalazłem w rozprawie jawnego sformułowania tzw. tezy doktorskiej czy hipotezy badawczej. Niespełnienie powyższego (zwyczajowego) wymogu jest w rozprawie zastąpione jednoznacznym określeniem celu rozprawy, którym jest zbudowanie wspomnianego powyżej algorytmu estymacji gęstości rozkładu danych strumieniowych. Można zatem przyjąć, że hipotezę badawczą stanowi

weryfikacja możliwości budowy tego rodzaju algorytmu o określonym poziomie skuteczności dla syntetycznych oraz rzeczywistych danych strumieniowych.

Treść rozprawy

Rozprawa liczy 76 stron, składa się z przedmowy, 7 rozdziałów oraz spisu literatury zawierającego 79 pozycji.

W przedmowie Doktorant uzasadnia znaczenie podjętej tematyki badawczej oraz przedstawia ramowy opis zawartości poszczególnych rozdziałów oraz wykaz osiągnięć publikacyjnych Doktoranta związanych bezpośrednio z badaniami opisanymi w dysertacji. Nie mam istotnych uwag do tego fragmentu rozprawy. Dostrzegłem jedynie kilka błędów literowych - w tym jeden o istotniejszym znaczeniu (zamiana potęgi E21 na E12 w oszacowaniu liczby przechowywanych danych).

W rozdziale pierwszym Autor przedstawia cel, zakres oraz motywację badań opisanych w rozprawie wskazując, że w przypadku danych strumieniowych konieczne jest dokonywanie istotnych modyfikacji klasycznych metod estymacji gęstości rozkładu bądź stosowanie nowych (dedykowanych danym strumieniowym) algorytmów. W szczególności, w przypadku niestacjonarnych strumieni danych konieczne jest wprowadzenie mechanizmów ważących istotność poszczególnych danych w kontekście momentu ich pojawienia się w analizowanym strumieniu.

Zagadnienie badawcze rozważane przez Kandydata ma bez wątpienia istotne znaczenie w kontekście lawinowo rosnącego wolumenu danych opisujących różnorodne procesy i zjawiska (np. w diagnostyce medycznej, analizach finansowych, energetyce, czy zagadnieniach bezpieczeństwa). Wiele spośród tych danych ma postać strumieniową o zmiennej w czasie charakterystyce.

Proponowane przez Autora podejście bazuje na powszechnie wykorzystywanej w ujęciu „klasycznym" technice estymatorów jądrowych, czyli metodzie nieparametrycznej estymacji gęstości rozkładu. Zmodyfikowana przez Autora wersja metody (o akronimie DEDSTA) wykorzystuje techniki prognozowania statystycznego do wykrycia niestacjonarności danych, bazuje na automatycznym doborze szerokości okna zawierającego aktualnie rozważane obserwacje oraz wykrywaniu i odpowiedniej obsłudze elementów nietypowych (nowopowstałych oraz wygasających).

Rozdział napisany jest poprawnie, choć nieco chaotycznie. Moim zdaniem należałoby go rozbudować, jednocześnie porządkując poszczególne wątki. W tym miejscu chciałbym poczynić uwagę natury bardziej ogólnej, dotyczącą różnych fragmentów rozprawy. Otóż sposób opisu zjawisk czy metod badawczych prezentowanych w rozprawie (zarówno autorskich jak i cytowanych z literatury) ma generalnie lakoniczny charakter. O ile trzeba oddać Autorowi, że każde z zagadnień opisane jest w sposób wystarczający do ich podstawowego zrozumienia, o tyle w wielu miejscach (np. w przeglądzie literatury przedstawionym w rozdziale 1, opisie metody wykrywania niestacjonarności strumienia danych w rozdziale 3.4, czy opisie procedury scalania starych elementów sekwencji omawianej w rozdziale 5) rozszerzenie opisu niewątpliwie podniosłoby czytelność rozprawy oraz pozytywnie wpłynęło na całościową ocenę dysertacji.

Podejście Doktoranta dość dobrze charakteryzuje zdanie zamieszczone na końcu strony 9 w omawianym rozdziale wstępnym: *„Proponowana koncepcja ma charakter kompleksowy i nie wymaga dogłębnych teoretycznych dywagacji ani żmudnych badań"*. Z jednej strony Autor ma rację, w tym sensie, że metoda wydaje się działać prawidłowo (przynajmniej w zakresie eksperymentalnej

weryfikacji przeprowadzonej przez Doktoranta). Z drugiej jednak strony, istotą badań naukowych jest właśnie analiza i wyjaśnianie zjawisk i procesów, badanie ich cech i właściwości w celu lepszego ich zrozumienia, modelowania i wykorzystania.

W rozdziale drugim przedstawiony jest przegląd literatury, podzielony na dwie części. Pierwsza z nich obejmuje nieparametryczne metody gęstości rozkładu stosowane w przypadku danych nie posiadających strumieniowej charakterystyki, ze szczególnym uwzględnieniem estymatorów jądrowych stanowiących podstawę rozważań prowadzonych w dysertacji.

Druga część omawia metody estymacji gęstości stosowane w przypadku danych strumieniowych, w tym metody oparte o estymatory jądrowe, sieci samoorganizujące się, analizę skupień oraz metody falkowe.

Przegląd literatury zajmuje mniej niż 1,5 strony i ma charakter czysto bibliograficzny. Cytowanych w nim jest ok. 30-40 prac pogrupowanych w ramach kilku kategorii, w zasadzie bez merytorycznego komentarza. Zdecydowanie zabrakło w tym miejscu pogłębionego opisu wymienianych metod.

Rozdział trzeci omawia podstawowe pojęcia wykorzystywane w rozprawie, począwszy od zdefiniowania pojęcia strumienia danych, omówienia zjawiska niestacjonarności (często określanego w literaturze jako *concept drift*) oraz wskazania czterech podstawowych typów dryftu: nagłego, przyrostowego, nawracającego oraz stopniowego.

Następnie wprowadzane są podstawowy estymatorów jądrowych oraz omawiane jest zagadnienie detekcji elementów nietypowych. Kolejne dwa podrozdziały dotyczą kluczowych z punktu widzenia pracy zagadnień: wykrywania niestacjonarności strumienia (rozdział 3.4) oraz prognozowania statystycznego w oparciu o wygładzanie wykładnicze (rozdział 3.5).

Rozdział 3.4 stanowi kolejny przykład „nadmiernej lakoniczności" Autora, który – po podaniu podstawowych informacji dotyczących testu niestacjonarności KPSS - rzuca czytelnika na głęboką wodę zdaniem *„Zdefiniujmy zatem następującą wielkość* sgmKPSS(KPSS) = sgm(0.995 KPSS – 2.932)", nie uzasadniając dlaczego liczymy tę wartość dla parametru 0.995 KPSS oraz dlaczego pomniejszamy ten parametr o 2.932. Oczywiście odpowiedź można znaleźć w cytowanych przez Autora pracach, ale wyjaśnienie w rozprawie tak podstawowego wzoru (z punktu widzenia prezentowanych rozważań) jest absolutnie niezbędne.

Kolejny przykład niedostatecznych wyjaśnień omawianych zagadnień znajduje się na następnej stronie. Autor pisze, że *„Wybór owej krotności został dokonany na drodze heurystycznej, zainspirowany metodą Zieglera-Nicholsa, podstawową w klasycznej teorii regulacji z zakresu automatyki, używanej do wyznaczania nastaw sterowników PID [9].*" Nie jestem automatykiem, nie wiem jak się wyznacza nastawy sterowników PID, nie znam także metody Zieglera-Nicholsa i myślę, że nie jestem w tym odosobniony. Z pewnością kilka zdań tłumaczących istotę wskazanego podejścia heurystycznego byłoby pomocne w zrozumieniu tego fragmentu pracy.

W rozdziale 3.5 omawiającym metodę prognozowania statystycznego wektor A_t ma w równaniu (3.40) ma postać wierszową, a w równaniu (3.42) kolumnową.

Kolejny rozdział przestawia główny wynik pracy czyli metodę DEDSTA (*Density Estimation for Data Stream with possible Trend Algorithm*). Idea metody polega na zastosowaniu ruchomego

okna o zmiennej w czasie szerokości, dobieranej w sposób adaptacyjny, próbkującego strumień danych. Kwestia doboru szerokości okna dyskutowana jest w rozdziale 4.1, następnie w rozdziale 4.2 omawiane jest zagadnienie dezaktualizacji danych w miarę upływu czasu, w tempie zależnym od stopnia zmienności strumienia. Zagadnienie określenia wspomnianego wyżej stopnia niestacjonarności strumienia danych rozważane jest w rozdziale 4.3. Kolejny punkt (4.4) odnosi się do problemu wykrywania obserwacji nietypowych. Rozdział kończy dyskusja dotycząca holistycznego spojrzenia na trzy wspomniane wyżej czynniki (dezaktualizacja danych, predykcja niestacjonarności oraz wykrywanie obserwacji nietypowych) w kontekście określenia finalnych wag przypisanych poszczególnym obserwacjom w metodzie DEDSTA.

Rozdział napisany jest starannie, zawiera wszystkie informacje niezbędne do zrozumienia omawianych treści, a struktura opisu – w odróżnieniu od poprzednich rozdziałów – jest uporządkowana.

Jak wspomniałem powyżej rozdział ten, obok rozdziału 6 przedstawiającego wyniki analizy eksperymentalnej metody jest kluczowym fragmentem dysertacji. Mam trzy uwagi dotyczące tego fragmentu pracy:

(1) Odnosząc się do wzoru (4.13) Autor pisze, że „*Ostatecznie proponowana jest* $\beta_0=2/3$". Dlaczego akurat taka wartość została przyjęta, skoro – zgodnie z zapisami akapitu powyżej - uznawana jest ona za graniczną? Naturalnym kandydatem wydaje się być wartość pośrednia pomiędzy 1/3 a 2/3, np. 1/2.

(2) We wzorze (4.12) pojawia się symbol m_t, który – jak sądzę – oznacza wartość m w chwili t, ale nie jest to w dysertacji wyjaśnione.

(3) Istotne wątpliwości budzi sposób ustalania ostatecznej wartości wagi w_i stanowiący przemnożenie trzech wag w_i^*, w_i^{**} oraz w_i^{***} odnoszących się odpowiednio do dezaktualizacji obserwacji, predykcji niestacjonarności oraz wykrywania obserwacji nietypowych. Przykładowo, układ wag $[w_i^*, w_i^{**}, w_i^{***}]=[1/2,1/2,2]$ ma taki sam skutek jak $[w_i^*, w_i^{**}, w_i^{***}]=[1,1,1/2]$. O ile w pierwszym przypadku, dwie pierwsze wagi oznaczają odpowiednio częściową dezaktualizację obserwacji oraz brak istotnej niestacjonarności, natomiast trzecia oznacza wykrycie elementu nietypowego związanego z nowym trendem, o tyle w zestawie drugim sytuacja jest istotnie różna: dwie pierwsze wagi maja znaczenie neutralne (np. nie występują reprezentowane przez nie zjawiska), a waga trzecia wskazuje na element nietypowy zanikającego trendu. Innym przykładem jest zestaw wag [1/4, 2, 2], który jest wynikowo równoważny, np. zestawowi neutralnemu [1,1,1]. Przykładów tego rodzaju jest oczywiście więcej.

Rozdział piąty przedstawia wybrane metody, z którymi porównywany jest algorytm DEDSTA. Kolejno są to: adaptacyjna metoda falkowa (rozdział 5.1), algorytm jąder klastrowych (rozdział 5.2) oraz sieci samoorganizujące się (rozdział 5.3). Motywacja Autora do wyboru powyższych metod wynika z powszechnego stosowania pierwszych dwóch w przypadku estymacji gęstości rozkładu danych stacjonarnych. Trzecia metoda (sieci Kohonena) jest strukturalnie zbliżona do metody jąder klastrowych, jednak jej realizacja bazuje na innym formalizmie - samoorganizacji.

Mam następujące uwagi do treści omawianego rozdziału.

(1) We wstępie Kandydat pisze, że „*Estymacja gęstości danych strumieniowych jest obecnie obiektem intensywnych badań, których wyniki stają się przedmiotem licznych publikacji w czasopismach naukowych.*" W tej sytuacji pewne zdziwienie budzi fakt, że wybrane do

porównania metody pochodzą odpowiednio z lat 2005 (metoda falkowa), 2007 (metoda jąder klastrowych) oraz 2012 (mapy Kohonena). Dlaczego Doktorant nie dokonał porównania z nowszymi, prawdopodobnie skuteczniejszymi metodami?

(2) Metoda falkowa, zastosowana w sposób bezpośredni, nie pozwala na uzyskanie zadowalających wyników. W związku z powyższym Kandydat dokonał różnorodnych jej modyfikacji, nie przeprowadzając jednakże szczegółowej weryfikacji ich skutków, np. analizy doboru parametrów wynikowego algorytmu. W konsekwencji trudno jest określić rzeczywistą skuteczność zmodyfikowanej metody falkowej.

(3) Na stronie 38 Autor wyjaśnia pojęcie jądra klastrowego następująco: „*Jest to obiekt zawierający średnią arytmetyczną obserwacji wchodzących w skład danego jądra klastrowego, ...*" co z oczywistych względów nie jest prawidłową definicją.

(4) Odnosząc się do metod wykorzystujących jądra klastrowe, z którymi DEDSTA jest porównywana, Autor pisze, że „*Opis metod przedstawionych w tych pracach zawierał ewidentne niedociągnięcia. Zostały one poprawione, co skutkowało polepszeniem jakości estymatora*". Podobnie jak w przypadku metody falkowej nie został przedstawiony ani precyzyjny opis wprowadzonych zmian, ani też nie wykazano, że metoda wynikowa, w zastosowaniu do strumieni danych, jest istotnie metodą silną. Ponownie zasadne jest pytanie, czy faktycznie nie ma metod estymacji rozkładu opublikowanych w ostatnich latach, dedykowanych strumieniom danych, z którymi metoda przedstawiona przez Doktoranta mogłaby być porównana?

(5) W dalszym opisie metody jąder klastrowych, odnosząc się do jej pierwotnej wersji, Kandydat pisze, że „*Autorzy proponują stała liczbę jąder klastrowych równą 100*" i jak rozumiem taką samą wartość przyjmuje w swoich badaniach. Z uwagi na istotną zmianę warunków stosowania metody (dane strumieniowe vs. dane wsadowe) należało przeprowadzić testy dotyczące wyboru optymalnej parametryzacji metody. Nie ma podstaw do stwierdzenia, że wartość 100 w przypadku danych strumieniowych (w szczególności niestacjonarnych) jest wartością optymalną czy nawet wystarczająco dobrą.

(6) Charakteryzując metodę sieci samoorganizujących się Autor stwierdza, że „*W ostatniej fazie wagi wszystkich neuronów wskazanych we wcześniejszych etapach treningu ulegają transformacji, aby zbliżyć ich wartość do wylosowanej próbki*". Takie masowe zmiany prowadziłyby natychmiast do niestabilności procesu uczenia sieci SOM. Zmianie ulegają jedynie neurony należące do topologicznego sąsiedztwa neuronu zwycięzcy. Zakładam, że taką sytuację Autor miał na myśli w powyższym zdaniu.

(7) Na stronie 43 Kandydat pisze, że „*Liczba neuronów w każdej sieci jest równa 100*". W tym miejscu można przywołać uwagę odnośnie konieczności doboru parametryzacji metod do nowych warunków ich stosowania. Liczba 100, która sprawdzała się w określonych poprzednich zastosowaniach nie koniecznie musi być wartością właściwą w zadaniu strumieniowym rozważnym przez Doktoranta.

(8) Na stronie 44, opisując metodę sieci samoorganizujących się, Autor pisze o „*scalaniu starych elementów sekwencji*" w oparciu o zmodyfikowaną dywergencję Kullbacka-Lieblera. Z zamieszczonego opisu nie potrafię wywnioskować na czym w istocie wspomniane scalanie polega oraz jaki jest związek pomiędzy wartościami kl_j^M a procesem scalania obserwacji.

Rozdział szósty poświęcony jest numerycznej weryfikacji skuteczności zaproponowanej metody DEDSTA, zaczynając od analizy efektywności metody na przykładzie danych syntetycznych – jednowymiarowego szeregu czasowego o zmiennej w czasie charakterystyce, z dodaną składową losową. Wyniki metody są porównywane z rezultatami uzyskanymi przez trzy podejścia konkurencyjne przedstawione w poprzednim rozdziale jednoznacznie wykazując przewagę zaproponowanego przez Doktoranta algorytmu nad rozważanymi konkurentami. „Łyżką dziegciu" jest wspomniany przeze mnie brak ustalenia realnej siły metod, z którymi algorytm DEDSTA jest porównywany.

W dalszej części rozdziału metoda DEDSTA stosowana jest do estymacji gęstości rozkładu strumieni danych meteorologicznych w oparciu o całoroczne pomiary temperatury (a w części eksperymentów również wilgotności) wykonane w trzech odmiennych klimatycznie miejscach: Minneapolis, Rio de Janeiro i Krakowie. Podobnie jak w przypadku danych syntetycznych skuteczność metody została potwierdzona zarówno w kontekście danych jednowymiarowych jak i wielowymiarowych (łączny rozkład temperatury oraz wilgotności w Krakowie).

Omawiany rozdział napisany jest bardzo dobrze. Autor wyciąga trafne wnioski z przeprowadzonych eksperymentów. Uzyskane wyniki potwierdzają skuteczność metody DEDSTA w odniesieniu do rozważanych w dysertacji strumieni danych.

Rozdział siódmy stanowi krótkie podsumowanie dysertacji. Autor przypomina motywację leżącą u podstaw prowadzonych badań, streszcza najistotniejsze cechy wprowadzonej przez siebie metody oraz wskazuje możliwe kierunki rozwoju przedstawionych w dysertacji badań. Rozprawę dopełnia spis literatury.

Oryginalny wkład Autora rozprawy

Oryginalny wynik badawczy Doktoranta dotyczy opracowania, implementacji oraz eksperymentalnej weryfikacji autorskiej metody estymacji gęstości rozkładu danych strumieniowych z wykorzystaniem ruchomego okna o szerokości automatycznie dopasowywanej do bieżącej charakterystyki strumienia. Dodatkowo, zaproponowana metoda wykorzystuje mechanizm dezaktualizacji danych, procedurę wykrywania niestacjonarności strumienia oraz algorytm wykrywania obserwacji nietypowych (zarówno nowopowstałych jak i zanikających).

Przedstawione w rozprawie wyniki badawcze zostały częściowo opisane w publikacji konferencyjnej (140 pkt na liście MEiN) oraz pracy złożonej do recenzji w czasopiśmie (brak szczegółowych danych).

Konkluzja

Przedstawiona do recenzji rozprawa zawiera oryginalne wyniki prac badawczych Doktoranta w obszarze metod estymacji gęstości rozkładu w przypadku danych strumieniowych (stacjonarnych bądź niestacjonarnych).

Praca napisana jest starannie, choć miejscami zdecydowanie zbyt lakonicznie, warstwa językowa jest na właściwym poziomie, a nieliczne błędy językowe czy interpunkcyjne, które zauważyłem mieszczą się w dopuszczalnym zakresie. Nie dostrzegłem w rozprawie istotnych

nieprawidłowości, a wymienione w recenzji uwagi nie podważają mojej ogólnie pozytywnej oceny dysertacji.

Rozprawa dotyczy aktualnej tematyki badawczej a jej treść dowodzi posiadania przez Autora wiedzy w zakresie rozważanych zagadnień. Tematyka oraz treść rozprawy mieszczą się w obszarze dyscypliny *informatyka techniczna i telekomunikacja.*

Reasumując, **stwierdzam, że rozprawa spełnia wymagania stawiane przez odnośną Ustawę i wnoszę o jej przyjęcie oraz dopuszczenie jej Autora, mgr inż. Tomasza Rybotyckiego, do dalszych etapów przewodu doktorskiego.**