dr hab. inż. Rafał Kozik, profesor uczelni
Politechnika Bydgoska
im. Jana i Jędrzeja Śniadeckich w Bydgoszczy,
Wydział Telekomunikacji, Informatyki i Elektrotechniki,
Al. prof. S. Kaliskiego 7,
85-796 Bydgoszcz

Bydgoszcz, 16.02.2022

RECENZJA ROZPRAWY DOKTORSKIEJ WYKONANA DLA RADY DYSCYPLINY NAUKOWEJ INFORMATYKA TECHNICZNA I TELEKOMUNIKACJA POLITECHNIKI WROCŁAWSKIEJ

Tytuł rozprawy: Zastosowanie zespołów klasyfikatorów w klasyfikacji niezbalansowanych strumieni danych

Autor rozprawy: mgr inż. Jakub Klikowski

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, itd.)?

Zakres rozprawy, której Autorem jest Pan mgr inż. Jakub Klikowski, dotyczy wykorzystania zespołów klasyfikatorów w binarnej klasyfikacji strumieni danych, które charakteryzują się niezbalansowanym rozkładem etykiet oraz wykazują dryft koncepcji.

Jest to aktualna i ważna tematyka. Jest wiele dziedzin nauki, gdzie zagadnienie binarnej klasyfikacji danych jest bardzo istotne (np. bezpieczeństwo sieciowe, szeroko pojęta diagnostyka, sprzedaż/marketing, itd.). W większości przypadków mamy do czynienia ze znaczącą i ciągle rosnącą ilością danych, których specyfikacja oraz rozkład cech jest dynamicznie zmienny. Dlatego poszukiwanie rozwiązań, które pozwalają osiągnąć coraz to lepszą jakość predykcji, jest jak najbardziej celowe.

W rozdziale pierwszym, Autor rozprawy precyzyjnie wskazał kluczowe wyzwania i problemy związane z binarną klasyfikacją niezbalansowanych strumieni danych, które posiadają dryft koncepcji.

W tym świetle zostaje jasno i precyzyjnie postawiona hipoteza badawcza rozprawy. Zakłada ona, że można zaprojektować takie zespoły klasyfikatorów (wykorzystujące próbkowanie danych oraz jednoklasowe klasyfikatory), które potrafią osiągnąć jakość predykcji lepszą niż znane z literatury klasyfikatory niezbalansowanych strumieni danych.

Rozprawa ma charakter teoretyczno-eksperymentalny. Autor definiuje kilka celów badawczych, które trafnie wpisują się w hipotezę rozprawy. Fakt zrealizowania każdego z nich, zostaje zweryfikowany eksperymentalnie. Każdy eksperyment jest odpowiednio zaplanowany a proponowane rozwiązanie zostało porównane z konkurencyjnymi rozwiązaniami, które znane są z literatury.

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?

Bibliografia niniejszej rozprawy zbudowana jest z ponad 300 pozycji. Podał połowa z nich została opublikowana w latach 2010-2021. Spora część dotyczy ostatnich pięciu lat. W szczególności jest dużo (według moich szacunków, ponad 60) publikacji, które datowane są na lata 2020, 2019 i 2018.

Analiza literaturowa wykonana jest zasadniczo w dwóch etapach. Ogólny zarys stanu wiedzy i problematyki związanej z niniejszą rozprawą przedstawiony został w rozdziale pierwszym. Liczne odwołania do literatury, które znajdują się w tym rozdziale, pozwoliły uwypuklić istotność takich problemów jak niezbalansowanie danych, big data, binarna klasyfikacja niestacjonarnych strumieni danych, czy dryft koncepcji. W tym rozdziale również znajduje się uzasadnienie zastosowania, przez Autora rozprawy, komitetu klasyfikatorów do kategoryzowania trudnych zbiorów danych. W szczególności, Autor wskazuje tutaj przykłady praktycznych aplikacji tego typu rozwiązania do detekcji spamu, rozpoznawania twarzy, kategoryzowaniu tekstu, czy diagnostyce.

Drugi i kluczowy etap analizy znajduje w rozdziale drugim. Poza ogólnym wprowadzeniem do zagadnień związanych z uczeniem maszynowym, klasyfikacją, komitetem klasyfikatorów oraz klasyfikacją strumieni, Autor rozprawy przedstawia konkurencyjne rozwiązania, które zostają szczegółowo porównane z autorskimi metodami w rozdziale dotyczącym eksperymentów. W szczególności są to algorytmy KMC, L++CDS, L++ NIE, REA, OUSE oraz MLPC.

W mojej ocenie wnioski z analizy literatury są sformułowane w sposób jasny i są trafne. W szczególności popieram stwierdzenie, iż niewiele jest prac zajmujących się problemem

klasyfikacji strumieni danych, które mają nierównomierny rozkład klas. Jak Autor rozprawy zauważył, często stosowaną techniką przy rozwiązywaniu problemu niezbalansowanych danych jest generowanie sztucznych próbek. Dlatego, słusznym i unikalnym wydaje się być, zaproponowany przez Autora, mechanizm akumulowania danych.

3. Czy Autor rozwiązał postawione zagadnienia? Czy użył do tego właściwych metod i czy przyjęte założenia są uzasadnione?

W mojej ocenie Autor w sposób właściwy rozwiązał postawiony problem. W procesie tym użyte zostały właściwe narzędzia i metody. Tym samym Autor dowiódł, że posiadł umiejętności związane z metodyką i metodologią prowadzenia badań. W szczególności badania i eksperymenty osadzone zostały na obiektywnej analizie literaturowej.

Autor rozprawy odniósł się także do aktualnych trendów i ograniczeń metod typowo wykorzystywanych przy klasyfikacji niezbalansowanych strumieni danych. W tym świetle zaprojektowane, zaimplementowane i przebadane zostały różne autorskie algorytmy i rozwiązania. W mojej ocenie, zaplanowane i przeprowadzone eksperymenty są wiarygodnym dowodem zrealizowania przez Autora celów badawczych. Na podkreślenie zasługuje fakt, iż autor rozprawy zdefiniował aż sześć celów badawczych.

Dodatkowo, udokumentowane w rozprawie wyniki świadczą o efektywności proponowanego rozwiązania w analizowanych scenariuszach badawczych. Autor rozprawy pokazał, iż proponowane podejścia dorównują lub przewyższają znane z literatury algorytmy. Warto podkreślić, iż niemal każde proponowane rozwiązanie daje inny wymiar wartości dodanej w stosunku do konkurencyjnych rozwiązań. Przykładowo, OCEIS daje zadowalające wyniki na rzeczywistych strumieniach danych, OCWE pozwala osiągnąć lepszą wydajność, itd.

Na podkreślenie, zasługuje również fakt, iż Autor rozprawy dokonał statystycznej analizy wyników otrzymanych z poszczególnych eksperymentów. Pozwala to jednoznacznie stwierdzić, że otrzymane wyniki są faktycznie statystycznie lepsze od tych, które zostały otrzymane z wykorzystaniem metod znanych z literatury.

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 reprezentowanej przez literaturę światową?

Oryginalność rozprawy oraz samodzielny dorobek Autora, stanowi propozycja wykorzystania hybrydowych technik próbkowania z akumulacją danych i modelami klasyfikatorów jednoklasowych. W szczególności w mojej ocenie oryginalnymi i unikatowymi elementami rozprawy są:

- Zespół klasyfikatorów jednoklasowych (OCEIS) zaproponowany w celu klasyfikacji niezbalansowanych strumieni danych
- Opracowanie zespołu klasyfikator o nazwie OCWE, który jest ulepszoną wersją OCEIS
- Opracowanie klasyfikatora DSC, który wykorzystuje hybrydową technikę próbkowania i akumulacji danych
- Wykorzystanie baggingu w celu ulepszenia klasyfikatora DSC
- Opracowanie klasyfikatora DSE, który dodaje mechanizm detekcji dryftu
- Opracowanie metody DSE-OC, która łączy zalety algorytmów DSE oraz OCEIS

W mojej ocenie, pozycja pracy jest w pełni zgodna z aktualnym stanem wiedzy oraz poziomem techniki reprezentowanej przez literaturę światową. Wysoko punktowane publikacje, w których Autor niniejszej rozprawy jest pierwszym autorem, pozwalają stwierdzić, iż zademonstrowane osiągnięcia naukowe stanowią znaczący wkład w literaturze międzynarodowej.

5. Czy Autor wykazał umiejętności poprawnego i przekonywującego przedstawienia uzyskanych przez siebie wyników (zwięzłość, jasność, poprawność redakcyjna rozprawy)?

Nie mam wątpliwości, że Autor rozprawy posiada dużą wiedzę dotyczącą zagadnień związanych z tematem rozprawy doktorskiej. Doktorant wykazał się przede wszystkim umiejętnościami dotyczącymi prowadzenia badań naukowych, eksperymentów oraz projektowania algorytmów.

Również zrealizowane przez Autora rozprawy eksperymenty są zgodne z dobrymi praktykami w zakresie pomiaru skuteczności systemów klasyfikacji. Autor w tym zakresie użył znanych metryk co pozwala odnieść się do innych rozwiązań dostępnych w literaturze.

6. Jakie są wady i słabe strony rozprawy?

Rolą recenzenta jest zauważenie i wskazanie pewnych niedociągnięć (czy mankamentów) ocenianej pracy, oraz zgłoszenie tych uwag, które mogą być potencjalnie przydatne w dalszych pracach badawczo-rozwojowych. Dlatego też do wad i słabych stron rozprawy zaliczam:

- Pominięcie w eksperymentach problemu wielowymiarowości. Tabele pokazujące konfiguracje generatora strumieni zazwyczaj zawierają informację o 10 cechach, gdzie 2 z nich są redundantne. Oczywiście ilość wymiarów w wektorze cech będzie ściśle powiązana z problemem, który chcemy rozwiązań. Przykładowo, przy klasyfikacji tekstu (np. analiza sentymentu) już pojedyncze słowo może być opisane wektorem, który ma kilkaset elementów. W przypadku całego zdania, liczba cech będzie znacząco wyższa. Uważam, że ciekawym elementem rozprawy byłoby stworzenie odpowiedni

wytycznych co do narzędzi oraz mechanizmów, które mogą być użyte w celu poradzenia sobie z problemem wielowymiarowości.

- O ile złożoność obliczeniowa jest precyzyjnie określona dla poszczególnych algorytmów, o tyle problem skalowalności mógłby być bardziej uwydatniony w rozprawie. W szczególności przydatnym elementem pracy mogłaby być analiza wpływu pewnych części algorytmów na możliwość replikowania proponowanych klasyfikatorów (albo zespołu klasyfikatorów) w obrębie strumienia. Przykładowo, czy możliwe jest uruchomienie N kopii zespołu klasyfikatorów, aby przyspieszyć przetwarzanie strumienia. Również, ciekawym zagadnieniem jest to czy zaproponowana w rozdziale 4.1 rezerwuar danych (data storage) może w skrajnych przypadkach okazać się wąskim gardłem.
- Niektóre z wykresów (np. 3.3 albo 3.4) mogłyby mieć większą czcionkę dla osi X. W wersji drukowanej odczytanie wartości krytycznej jest trudne.
- W niektórych przypadkach, wykresy radarowe (np. 3.9 b) mogłyby mieć węższy zakres wartości (np. 0.5 do 1.0). Wówczas różnice pomiędzy metodami mogłyby być bardziej dostrzegalne.

W pracy można dopatrzeć się kilku, mało istotnych błędów edytorskich:

- brak nawiasu zamykającego po Tab.3.1
- pozostawianie przypisu dolnego na stronie 64, gdy odsyłacz znajduje się na stronie 63
- pozycja literaturowa [67] nie posiada roku publikacji

7. Jaka jest przydatność rozprawy dla nauk technicznych?

Pomimo przedstawionych powyżej uwag, rozprawa mgr inż. Jakuba Klikowskiego posiada wiele silnych stron. Przede wszystkim rozprawa dotyczy ważnej tematyki i wnosi interesujący wkład do zagadnień związanych z klasyfikacją niezbalansowanych strumieni danych. W szczególności, zawiera ona autorską propozycję metod wykorzystujących hybrydowe techniki próbkowania z akumulacją danych. Jak pokazały wyniki z eksperymentów, proponowane rozwiązanie można wykorzystać do różnego rodzaj strumieni danych (np. rzeczywistych, syntetycznych, z różnego rodzaju dryfem koncepcji, itd.).

Warto także zauważyć, że Autor ma udokumentowany dorobek publikacyjny w zakresie rozprawy. W publikacjach wysoko punktowanych Doktorant jest pierwszym autorem, więc jego wkład można uznać za znaczący.

Wniosek

Niewątpliwie recenzowana rozprawa doktorska dowodzi dużej wiedzy Autora dotyczącej zagadnień związanych z rozprawą doktorską. Liczne i rozbudowane eksperymenty dowodzą skrupulatności i pozwalają stwierdzić, że Autor rozprawy opanował technikę planowania i prowadzenie badań naukowych. Ponadto, recenzowana praca jasno formułuje tezę, która została udowodniona poprzez badania eksperymentalne i realizację wszystkich celów badawczych.

Wobec powyższego stwierdzam, że recenzowana praca **spełnia wymagania stawiane rozprawom doktorskim** przez obowiązujące przepisy. Dlatego wnoszę o przyjęcie niniejszej rozprawy i **dopuszczenie mgr inż. Jakuba Klikowskiego do publicznej obrony.**