Bydgoszcz, 18.06.2021r.

dr hab. inż. Tomasz Andrysiak
Profesor uczelni
Wydział Telekomunikacji, Informatyki i Elektrotechniki
Uniwersytet Technologiczno-Przyrodniczy w Bydgoszczy
e-mail: andrys@utp.edu.pl
tel: +48 602 605 500

RECENZJA ROZPRAWY DOKTORSKIEJ

dla Rady Dyscypliny Naukowej Informatyka Techniczna i Telekomunikacja

Politechniki Wrocławskiej

Niniejsza recenzja została przygotowana w wyniku powołania na recenzenta rozprawy doktorskiej pt. „*Classifier selection for imbalanced data stream classification*" przez Radę Dyscypliny Naukowej - Informatyka Techniczna i Telekomunikacja na posiedzeniu w dniu 21 kwietnia 2021 roku oraz na podstawie pisma Prorektora Politechniki Wrocławskiej prof. Andrzeja Ożyhara z dnia 23.04.2021 roku.

Tytuł rozprawy: Classifier selection for imbalanced data stream classification.

Autor rozprawy: mgr inż. Paweł Zyblewski.

Promotor rozprawy: Prof. dr hab. inż. Michał Woźniak.

Podstawa opracowania recenzji:

- otrzymany egzemplarz rozprawy doktorskiej,
- pismo Przewodniczącego Rady Naukowej Dyscypliny „Informatyka Techniczna i Telekomunikacja" Politechniki Wrocławskiej prof. dr hab. inż. Michała Woźniaka z dnia 21.04.2021 roku,
- zawiadomienie o wyznaczeniu na recenzenta w postępowaniu o nadanie stopnia doktora z dnia 23.04.2021r.

Zawartość treściowa rozprawy oraz analiza zagadnienia naukowego

Recenzowana rozprawa doktorska dotyczy zagadnień związanych z wykorzystaniem algorytmów dynamicznej selekcji zespołu klasyfikatorów w połączeniu z metodami przetwarzania wstępnego w celu klasyfikacji statycznych oraz strumieniowych danych niezbalansowanych. Głównym celem pracy było przedstawienie naturalnej zdolności algorytmów selekcji klasyfikatorów do radzenia sobie z

niezbalansowaniem danych oraz zaproponowanie nowych, efektywnych rozwiązań rzadko poruszanego w literaturze problemu klasyfikacji wysoce niezbalansowanych strumieni danych.

Praca składa się z sześciu rozdziałów, bibliografii oraz streszczenia jak również wykazu używanych skrótów i symboli. W rozdziale pierwszym przedstawiono wprowadzenie do tematyki rozprawy zawierające motywacje oraz wyzwania badawcze, zarysowano cele i zakres pracy jak również sformułowano hipotezę badawczą. Rozdział drugi zawiera omówienie obszarów, które stanowią podstawę recenzowanej rozprawy i są niezbędne do prawidłowego wyjaśnienia proponowanych pomysłów. W pierwszej kolejności zostały przedstawione podstawy rozpoznawania wzorców, w tym sformułowano zadanie klasyfikacji, przedstawiono wprowadzenie do zespołu klasyfikatorów, a także zarysowano pojęcie różnorodności oraz problem doboru klasyfikatora. Omówiono również podejścia do oceny klasyfikatorów dla danych niezbalansowanych oraz strumieniowych jak również zaprezentowano opracowany w Python-ie pakiet do analizy strumieni danych. W trzecim rozdziale zaprezentowane zostały metody dedykowane zadaniu trudnej klasyfikacji danych stacjonarnych. Przedstawiono trzy podejścia oparte na grupowaniu modeli bazowych w jednowymiarowej przestrzeni różnorodności klasyfikatorów. Zaproponowano również nowe kryterium informujące o wpływie poszczególnych klasyfikatorów na różnorodność osiąganą przez cały zespół. Rozdział czwarty skupia się na połączeniu dwóch ważnych tematów badawczych związanych z analizą danych, tj. klasyfikacji danych strumieniowych oraz analizie danych z niezrównoważonym rozkładem klas. Zaproponowano nowe algorytmy zaprojektowane specjalnie do tego rodzaju zadań, wykorzystujące metody dynamicznej selekcji klasyfikatorów. Przestawiono również dwa nowe podejścia integrujące metody wstępnego przetwarzania danych i dynamicznego wyboru zespołów do klasyfikacji niezrównoważonych strumieni danych. W pierwszym przypadku jako klasyfikatory bazowe użyto rozwiązań związanych z rozpoznawaniem pojedynczego wzorca, podczas gdy drugie podejście wykorzystywało stratyfikowaną wersję Bagging do generowania klasyfikatorów bazowych. Rozdział piaty opisuje nową metodę aktywnego uczenia się klasyfikatora dla danych strumieniowych. Zaproponowano tam strategie budowania modeli klasyfikacji w przypadku strumieni danych z ograniczonym dostępem do etykiet, która wykorzystuje losowe podejście do etykietowania z rozwiązaniami właściwymi dla algorytmów uczenia aktywnego. Ostatni rozdział zawiera wnioski oraz podsumowanie jak również kierunki dalszych prac rozwojowych.

W rozprawie postawiono następującą hipotezę badawczą: "*Istnieją metody wykorzystujące zarówno wstępne przetwarzanie danych, jak i metody selekcji klasyfikatorów, które przewyższają jakość predykcji znanych z literatury metod stosowanych w klasyfikacji danych trudnych*". Przedmiotowa hipoteza została sformułowane poprawnie oraz adekwatnie (w sposób jasny i rzeczowy) do postawionego celu głównego oraz celów szczegółowych rozprawy. A także została całkowicie udowodniona w rozprawie w oparciu o otrzymane wyników eksperymentów komputerowych.

Przydatność rozprawy dla gospodarki narodowej

Tematyka rozprawy mieści się w dyscyplinie naukowej Informatyka Techniczna i Telekomunikacja. Z punktu widzenia nauk technicznych przydatność rozprawy dotyczy przede wszystkim opracowania algorytmów w obszarze problemów klasyfikacji danych trudnych, jak również zaprojektowania oraz implementacji środowiska eksperymentalnego w języku Python pozwalającego na przeprowadzenie badań związanych z klasyfikacją danych niezbalansowanych. Należy podkreślić również iż wszystkie opracowane metody klasyfikacji zostały poprawnie zweryfikowane na drodze licznie przeprowadzonych eksperymentów komputerowych. Przyjęte przez Autora założenie rozprawy są uzasadnione i dotyczą z jednej strony aktualnych problemów badawczych, a z drugiej znajdują odzwierciedlenie w realnych potrzebach rozwiązywania problemów praktycznych. Tematyka rozprawy jest niezwykle interesująca jak i w pełni uzasadniona a także wykazuje duży potencjał wykorzystania w rzeczywistych problemach dotyczących klasyfikacji statycznych oraz strumieniowych danych niezbalansowanych.

Charakterystyka rozwiązania postawionego zagadnienia

Autor rozwiązał zagadnienie zdefiniowane w hipotezie rozprawy. W tym celu sformułował własne autorskie koncepcje dotyczące: wykorzystania algorytmów dynamicznej selekcji zespołu klasyfikatorów w kontekście metod i technik przetwarzania wstępnego w zadaniach klasyfikacji danych statycznych oraz strumieniowych w szczególności niezbalansowanych. Zastosowana metodologia badawcza sprowadza się do następującej sekwencji działań: zdefiniowania problemu badawczego, propozycji rozwiązania postawionego problemu, zdefiniowania środowiska eksperymentowania z wykorzystaniem adekwatnych do postawionego problemu metod i technik oraz analizy otrzymanych wyników z wykorzystaniem odpowiednich wskaźników jakości. Tego typu postepowanie jest typowym podejściem stosowanym przez społeczność naukową w analizie zagadnień związanych z uczeniem maszynowym, tak więc zaprezentowane w dysertacji podejście jest prawidłowe oraz przemyślne jak również właściwe w kontekście tematyki rozprawy.

Charakter rozprawy

Recenzowana praca ma charakter koncepcyjno-eksperymentalny. Autor zaproponował szereg rozwiązań koncentrujących się na efektywnym rozwiązaniu rzadko poruszanego w literaturze problemu klasyfikacji wysoce niezbalansowanych strumieni danych. Uzyskane wyniki potwierdzają w pełni postawioną na wstępie pracy hipotezę badawczą. Dysertacja zawiera niezbędne elementy eksperymentu komputerowego, które pozwalają na oceną jakości opracowanych metod klasyfikacji, a poruszana tematyka jest aktualna co do stanu wiedzy dotyczącego wykorzystania metod uczenia maszynowego w zagadnieniach klasyfikacji danych trudnych.

Oryginalność rozprawy

Wkład autora w rozwój metod i technik związanych z uczeniem maszynowym w kontekście dynamicznej selekcji zespołu klasyfikatorów w połączeniu z metodami przetwarzania wstępnego w celu klasyfikacji statycznych oraz strumieniowych danych niezbalansowanych polega na opracowaniu i implementacji:

- algorytmu selekcji zespołu klasyfikatorów wraz z dedykowaną reguła kombinacji dla danych niezbalansowanych, który zrealizowano poprzez opracowanie trzech rozwiązań opartych na grupowaniu modeli bazowych w jednowymiarowej przestrzeni różnorodności klasyfikatorów, gdzie podstawą utworzenia tej przestrzeni stanowiła zaproponowana miara H, informującą o wpływie poszczególnych klasyfikatorów na różnorodność osiąganą przez cały ich zespół,
- algorytmu dynamicznej selekcji klasyfikatorów w oparciu o miary dystansu dla danych niezbalansowanych, który zrealizowano poprzez opracowanie dwóch rozwiązań oceniających kompetencje modeli bazowych w zależności od decyzji podjętych przez nie w odniesieniu do przypadków znajdujących się w lokalnym sąsiedztwie klasyfikowanej instancji jednocześnie uwzględniając odległość Euklidesową do tych przypadków,
- algorytmu opartego o przetwarzanie wsadowe dla zadań klasyfikacji wysoce niezbalansowanych strumieni danych, który zrealizowano poprzez wykorzystanie rozwiązania dokonującego klasyfikacji z użyciem reguły decyzyjnej, która wykorzystuje lokalną charakterystykę danych dla preferowania klasy mniejszościowej,
- metody łączącej dynamiczną selekcję klasyfikatorów oraz przetwarzanie wstępne danych na potrzeby klasyfikacji niezbalansowanych danych strumieniowych, którą zrealizowano poprzez opracowanie dwóch rozwiązań tj. metody generującej pojedynczy model dla każdej nowej porcji danych oraz metody wykorzystującej do tego celu stratyfikowana wersję Bagging-u.

Ponadto do oryginalnych osiągnięć rozprawy można zaliczyć zaprojektowanie, implementację oraz ewaluację strategii budowania modeli klasyfikacji w przypadku strumieni danych z ograniczonym dostępem do etykiet.

W tym kontekście zaproponowano strategie odpytywania o etykiety łączącą losowe podejście do etykietowania z podejściem właściwym dla algorytmów uczenia aktywnego opartym na przekazywaniu do etykietowania przypadków znajdujących się w określonej odległości od granicy decyzyjnej problemu. Oryginalnym osiągnieciem recenzowanej rozprawy jest również wykazanie skuteczności opracowanych rozwiązań poprzez realizację ewaluacji eksperymentalnej porównującej zaproponowane algorytmy z podejściami opisanymi w literaturze przedmiotu i stanowiącymi state-of-the-art rozpatrywanego problemu. Zrealizowano to poprzez zaprojektowanie oraz implementację środowiska eksperymentalnego w języku Python, które wykorzystano do przeprowadzenia badań związanych z klasyfikacją danych niezbalansowanych. Ponadto Autor rozprawy opracował bibliotekę stream-learn, pozwalającą na przetwarzanie niezbalansowanych danych z dryftem koncepcji. Biblioteka ta została wykorzystana do przeprowadzenia wszystkich eksperymentów związanych z danymi strumieniowymi.

Ocena dorobku publikacyjnego doktoranta

W recenzowanej rozprawie doktorskiej wykazano znaczący dorobek publikacyjny doktoranta składający się z 16 wyselekcjonowanych pozycji (z uwzględnieniem 4 prac w procesie recenzji) w tym: 9 pozycji, których doktorant rozprawy jest pierwszym Autorem oraz 7 pozycji w których jest drugim Autorem. Przedmiotowe publikacje zostały zawarte w uznanych czasopismach międzynarodowych tj. Information Fusion (IF 13,669), czy też Pattern Analysis and Applications jak również w materiałach z konferencji międzynarodowych w większości rankingowanych jako CORE A tj. International Joint Conference on Neural Network, International Conference on Computational Science jak również Machine Learning and Knowledge Discovery in Databases. Oceniając dorobek publikacyjne doktoranta należy stwierdzić iż jest on znaczący, zważywszy, że dotyczy krótkiego okresu 2019-2021, w którym tylko jedna praca wydana była w 2019 roku. Należy również zauważyć, że wszystkie publikacje Aurora koncentrują się na problematyce uczenia maszynowego, w szczególności szeroko rozumianych zagadnieniach klasyfikacji statycznych oraz strumieniowych danych niezbalansowanych.

Dyskusyjne strony rozprawy

Pomimo osiągniecia postawionych celów rozprawy, udowodnienia sformułowanej hipotezy oraz poprawnej ewaluacji eksperymentalnej, w recenzowanej rozprawie można zauważyć pewne aspekty dyskusyjne, wymagające wyjaśnienia i doprecyzowania, jak również mniej znaczące usterki natury edycyjnej. Należy podkreślić i zaznaczyć iż recenzowana rozprawa nie zawiera słabych stron a jedynie pewne aspekty wymagające odniesienia w toku obrony rozprawy doktorskiej.

Uwagi mające charakter dyskusyjny można sformułować następująco:

- W pracy dość mało miejsca poświęcono problemowi ograniczonego dostępu do etykiet. Co prawda zaproponowano kilka algorytmów, w tym algorytm BALS (Budget Active Labeling Strategy), jednakże znane strategie uczenia przy braku dostępu do wszystkich etykiet wykorzystują inne mechanizmy uczenia aktywnego, np. algorytmy grupowania, ale także podejścia uczenia pół-nadzorowanego, gdzie wykorzystywane są do budowy modelu także obiekty nieetykietowane. Czy doktorant rozważał wykorzystanie innych strategii niż proponowanych w pracy, a także czy dostrzega jakieś ograniczenia dla stosowania tych metod w rozważanych przez niego zadaniach.
- W algorytmach pruning-u zespołów klasyfikatorów wykorzystano znane miary dywersyfikacji (diversity measure). Bazują one na wielkościach frakcji obiektów, na których klasyfikatory z ocenianej grupy podejmują różne decyzje. Czy doktorant zgodzi się z twierdzeniem, że wykorzystywane miary są dedykowane prostej metodzie głosowania większościowego, którą zresztą wykorzystano w badaniach? Jeżeli tak, to czy zmiana reguły uzyskania decyzji końcowej przez grupę predykatorów powinna być także uwzględniona w wykorzystywanej mierze dywersyfikacji?

- W pracy wykorzystano proste metody bagging-u, gdzie do generowania zbiorów perturbowanych wykorzystano proste losowanie ze zwracaniem (z reguły zapewniającym próbkę stratyfikowaną) – tzw. nieparametryczny bagging. Dlaczego doktorant zakłada rozkład jednostajny obiektów przy tworzeniu perturbowanych zbiorów, a nie stosuje innych wariantów bagging-u z innym niż jednostajny rozkład?

Recenzowana rozprawa napisana jest starannie pod względem językowym, stylistycznym oraz redakcyjnym. Niemniej jednak w pracy można znaleźć nieliczne błędy redakcyjne, m. in. takie jak:

- str. 14 brak kropka na końcu zdania „... a joint result",
- str. 24 brak dwukropka na końcu zdania „... in the classification task",
- str. 28 przecinek zamiast kropki na końcu równania 2.22,
- str. 59 kropki zamiast przecinaków w tekście u dołu strony,
- str. 157 brak kropki na końcu zdania „... instances each",
- str. 183-207 brak jednolitego standardu zapisu pozycji literaturowych oraz nieliczne błędy interpunkcyjne.

Należy podkreślić, że przytoczone powyżej błędy redakcyjne nie pomniejszają wartości naukowej oraz oryginalności rozprawy.

Ocena wiedzy Autora oraz znajomości współczesnej literatury z zakresu dyscypliny

Rozdziały nr 1 oraz 2 opisują kolejno: wyzwania badawcze w kontekście natury niezbalansowanych danych oraz stan wiedzy dotyczący uczenia maszynowego, w szczególności sformułowano tam zadanie klasyfikacji, przedstawiono koncepcję zespołu klasyfikatorów, a także zarysowano pojęcie różnorodności oraz problem wyboru odpowiedniego klasyfikatora. Przedstawione w wymienionych rozdziałach treści wskazują, że Autor rozprawy posiada obszerną wiedze teoretyczną, która dotyczy omawianej w pracy problematyki i mieści się w aktualnym nurcie badań związanych z uczeniem maszynowym, w szczególności z zagadnieniami klasyfikacji. Treści tych rozdziałów odnoszą się do aktualnego stanu wiedzy i odpowiednio wprowadzają czytelnika do poszczególnych zagadnień, które w dysertacji stanowią oryginalny wkład Autora. Spis literatury liczy 284 pozycje. Cytowane prace dobrane są prawidłowo i odnoszą się do omawianych w pracy problemów oraz świadczą o dużej umiejętności korzystania z istniejącej literatury przez doktoranta.

Ocena umiejętności poprawnego i przekonującego przedstawienia wyników

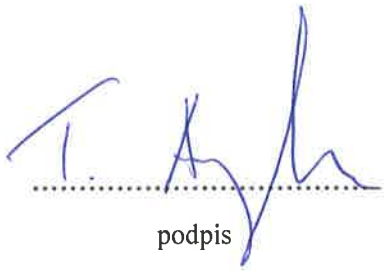
Autor wykazał umiejętności poprawnego i przekonującego przedstawienia wyników działania opracowanych autorskich algorytmów oraz ich eksperymentalnej weryfikacji z wykorzystaniem referencyjnych zbiorów danych a także własnego środowiska eksperymentalnego w języku Python oraz biblioteki stream-learn. Styl i forma prezentacji jest czytelna, poprawna i zrozumiała. Dodatkowo wyniki badań eksperymentalnych zostały szeroko skomentowane w treści rozprawy. Należy stwierdzić iż niewątpliwie Autor rozprawy posiada duże umiejętności związane z prezentowaniem oraz analizą wyników prac badawczych.

Podsumowanie

Przedstawiona do recenzji rozprawa doktorska mgr inż. Pawela Zyblewskiego pt. *„Classifier selection for imbalanced data stream classification"*, której promotorem jest Prof. dr hab. inż. Michał Woźniak, spełnia wszystkie wymagania stawiane rozprawom doktorskim przez *Ustawę z dn. 20 lipca 2018: Prawo o szkolnictwie wyższym i nauce (Dz. U. 2021, poz. 478).*

Recenzowana rozprawa doktorska stanowi samodzielne, bardzo dobre rozwiązanie problemu badawczego mieszczącego się w zakresie dyscypliny naukowej *informatyka techniczna i telekomunikacja* w dziedzinie nauk inżynieryjno-technicznych i potwierdza wysokie umiejętność rozwiązywania problemów naukowych zgodnych z jej zakresem przez Autora rozprawy. Na uznanie zasługuje również duża aktywność publikacyjną doktoranta.

Wnioskuje o przyjęcie recenzowanej rozprawy i dopuszczenie jej do publicznej obrony oraz uhonorowanie jej wyróżnieniem.



podpis