dr hab. inż. Robert Burduk
Politechnika Wrocławska
Wydział Elektroniki
Ul. Wybrzeże Stanisława Wyspiańskiego 27
50-370 Wrocław

Wrocław, dnia 16.08.2021 r.

RECENZJA

rozprawy doktorskiej mgra inż. Michała Koziarskiego
zatytułowanej: **„Imbalanced data preprocessing techniques utilizing local data characteristics"**

Recenzja została sporządzona w związku z powołaniem przez Radę Dyscypliny Informatyka Techniczna i Telekomunikacja Akademii Górniczo-Hutniczej im. Stanisława Staszica w Krakowie, piszącego niniejszą recenzję, jako recenzenta rozprawy doktorskiej mgra inż. Michała Koziarskiego pismem z dnia 15 czerwca 2021 r.

Kryteria oceny dysertacji wynikają z przepisów zawartych w art. 13 ust. 1 Ustawy z dnia 14 marca 2003 r. o stopniach naukowych i tytule naukowym oraz o stopniach i tytule w zakresie sztuki (Dz. U. nr 65 r., poz. 595 z późn. zm.).

Problem badawczy i jego znaczenie

Zakres recenzowanej rozprawy dotyczy szerokiej i dynamicznie rozwijającej się problematyki uczenia maszynowego. W recenzowanej rozprawie Doktorant koncentruje się na zagadnieniach dotyczących przetwarzania wstępnego tzw. danych niezbalansowanych, które charakteryzują się nierównomiernym prawdopodobieństwem *a priori* etykiet klas występujących w zbiorze uczącym. Problematyka poruszona w dysertacji jest jednym z istotnych nurtów uczenia maszynowego, który jest aktualny i znajduje odzwierciedlenie w problemach rzeczywistych, w których korzystając z dostępnych danych uczących nie można zagwarantować występowania równomiernej liczby obiektów z poszczególnych etykiet klas. Celem przetwarzania wstępnego danych niezbalansowanych jest modyfikacja liczby obiektów przypisanych do etykiety klasy mniejszościowej lub/i większościowej. Przeprowadzenie procesu przetwarzania wstępnego pozwala na zastosowanie tradycyjnych metod uczenia

maszynowego, które bez tego procesu preferują (są stronnicze) w przypisywaniu obiektów do etykiety klasy większościowej.

W rozprawie sformułowano hipotezę badawczą, która zakłada, że wykorzystanie w procesie przetwarzania wstępnego danych niezbalansowanych pewnych charakterystyk wynikających z położenia obiektów w przestrzeni cech może skutkować lepszą jakością klasyfikacji algorytmów uczenia maszynowego. W szczególności porównanie wyników klasyfikacji dotyczy wykorzystania w procesie rozpoznawania innych, znanych z literatury, metod przetwarzania wstępnego danych niezbalansowanych. Postawiona hipoteza badawcza nie budzi zastrzeżeń i jednocześnie definiuje cel badawczy dysertacji, którym jest porównanie proponowanych przez Doktoranta metod przetwarzania wstępnego danych niezbalansowanych z metodami znanymi z literatury.

Eksperymentalna weryfikacja postawionej hipotezy badawczej została wykonana z wykorzystaniem wielu zbiorów danych pochodzących z publicznych repozytoriów danych oraz danych rzeczywistych dotyczących diagnostyki medycznej, w szczególności obrazów histopatologicznych.

Tematyka podjęta przez mgra inż. Michała Koziarskiego jest interesująca, w pełni uzasadniona i odpowiada na wyzwania współczesnej informatyki, które dotyczą między innymi coraz większej różnorodności, a tym samym trudności danych, które wykorzystywane są w uczeniu maszynowym. Recenzowana rozprawa bez wątpienia podejmuje wątek badawczy mieszczący się w zakresie dyscypliny Informatyka Techniczna i Telekomunikacja.

Struktura pracy oraz wiedza Autora

Recenzowana praca została napisana w języku angielskim i liczy 70 stron maszynopisu. Składa się z czterech rozdziałów merytorycznych, wstępu, zakończenia, bibliografii, streszczenia w języku polskim oraz angielskim. Rozdział 1 przedstawia kolejno: listę publikacji mgra inż. Michała Koziarskiego, motywację dotyczącą podjęcia tematyki badawczej, sformułowaną hipotezę badawczą, podsumowanie wkładu Doktoranta w problematykę przetwarzania wstępnego danych niezbalansowanych, strukturę dysertacji oraz podziękowania.

Rozdział 2 zawiera wprowadzenie do problematyki dotyczącej danych niezbalansowanych. W szczególności Autor przedstawił zagadnienia związane z modyfikowaniem liczby obiektów z poszczególnych etykiet klas (podpróbkowanie oraz nadpróbkowanie), omówił wady istniejących algorytmów zmiany liczby obiektów (w szczególności algorytm nadpróbkowania SMOTE). W Rozdziale 2 znajduje się również

przedstawienie problematyki dekompozycji niezbalansowanego zdania wieloklasowego na zadania binarne. Treść Rozdziału 2 jest przestawiona w sposób syntetyczny i wskazuje na problemy badawcze, których dotyczy rozprawa. Omawiany rozdział mógłby jednak wprowadzać Czytelnika w szerszy kontekst dotyczący problematyki danych niezbalansowanych np. wskazując na inny nurt badań, który dotyczy modyfikacji algorytmów uczenia maszynowego w celu redukcji preferencji etykiety klasy większościowej, czy też przedstawiając różne definicje wskaźnika niezbalansowania (skośności zbioru danych).

Rozdział 3 dysertacji przedstawia autorskie algorytmy przetwarzania wstępnego danych niezbalansowanych dedykowane do zadania klasyfikacji binarnej. Schemat przedstawienia algorytmów jest następujący: opis algorytmu zawierający motywacją lub słabe stronny innych metod pod- lub nadpróbkowania, wizualizację działania algorytmu z wykorzystaniem danych syntetycznych, kod algorytmu oraz odwołanie do najistotniejszych wyników badań eksperymentalnych. Opis badań, protokół eksperymentalny oraz szczegółowe wyniki zawarte są w cytowanych pracach mgra inż. Michała Koziarskiego. Pewnym mankamentem skrótowego przedstawienia wyników badań jest nie umieszczenie bezpośrednio w rozprawie wszystkich wniosków wynikających z przeprowadzonych badań eksperymentalnych lub osiągniętych rezultatów. Jako przykład może posłużyć opis złożoność obliczeniowej algorytmu z Rozdziału 4.2, który nie znajduje się bezpośrednio w dysertacji. Wzmianka o złożoności obliczeniowej znajduje się natomiast w przypadku algorytmu opisanego w Rozdziale 3.3.

W Rozdziale 4 Doktorant przestawił dwa autorskie algorytmy, których celem jest modyfikacja liczby obiektów w problemie wieloklasowym. Opracowane algorytmy wykorzystują wcześniej opisane metody nadpróbkowania, a struktura rozdziału jest taka sama jak Rozdziału 3.

Badania eksperymentalne wykonane na rzeczywistych zbiorach danych zostały przedstawione w Rozdziale 5. W opisanych badaniach wykorzystano dane dotyczące analizy obrazów histopatologicznych. Jest to publicznie dostępny zbiór danych: Breast Cancer Histopathological Image Classification (BreakHis) oraz zbiór danych o nazwie DiagSet, który powstał przy współpracy z Zakładem Opieki Zdrowotnej Diagnostyka Consilio Sp. z o.o w ramach realizacji grantu z Programu Operacyjnego Inteligentny Rozwój (POIR) finansowanego przez NCBiR.

Spis literatury liczy 72 pozycje. Cytowane prace dobrane są prawidłowo, są aktualne i odnoszą się do omawianych problemów.

Praca napisana jest bardzo starannie pod względem edycyjnym. Zamieszczenie spisu oznaczeń ułatwiłoby natomiast lekturę dysertacji.

Wkład Autora — oryginalne osiągnięcia

Wkład Autora w rozwój dyscypliny Informatyka Techniczna i Telekomunikacja dotyczy: zaproponowania (lub modyfikacji wcześniej opracowanych przez Doktorana) algorytmów przetwarzania wstępnego danych niezbalansowanych binarnych jak i wieloklasowych. Wkład utylitarny dotyczy natomiast wykorzystania zaproponowanych metod w systemie klasyfikacji obrazów histopatologicznych, który powstał w ramach projektu finansowanego prze NCBiR przy współpracy z otoczeniem gospodarczym.

Oryginalne osiągnięcia mgra inż. Miachała Koziarskiego przedstawione w dysertacji to:

1. Opracowanie algorytmu Combined Cleaning and Resamoling (CCR), który składa się z etapu czyszczenia (przesuwania obiektów należących do etykiety klasy większościowej) otoczenia obiektów należących do etykiety klasy mniejszościowej. Kolejny etap to nadpróbkowanie, które wykonywane jest w hipersferze, której centrum wyznaczone jest przez poszczególny obiekt należący do etykiety klasy mniejszościowej. Ta sama hipersfera (w sensie rozpatrywanego obiektu) wykorzystywana jest w etapie czyszczenia. W algorytmie CCR lokalna charakterystyka danych zdefiniowana jest zatem przez otoczenie hipersferyczne każdego obiektu należącego do etykiety klasy mniejszościowej.
2. Opracowanie algorytmów nadpróbkowania Radial-Based Oversampling (RBO) oraz podpróbkowania Radial-Based Undersampling (RBU), w których lokalna charakterystyka danych wykorzystywana jest do zdefiniowania gaussowskiej radialnej funkcji bazowej. W zaproponowanych algorytmach wartości tej funkcji wykorzystywane są do modyfikowania liczby obiektów przypisanych do etykiety klasy większościowej (RBU) lub etykiety klasy mniejszościowej (RBO). W każdym z tcyh algorytmów gaussowska radialna funkcja bazowa jest definiowana z wykorzystaniem innych obiektów, co wynika z charakterystyki opracowanych algorytmów.
3. Opracowanie algorytmu Radial-Based Combined Cleaning and Resampling (RB-CCR), który łączy pewne cechy wcześniej opracowanych algorytmów CCR oraz RBO. Hipersfera określana dla obiektu z etykiety klasy mniejszościowej dzielona jest arbitralnie na trzy regiony (o niskim, równym oraz wysokim potencjale), które zdefiniowane są przez wartości gaussowskiej radialnej funkcji bazowej. W poszczególnych regionach dokonywany jest proces nadpróbkowania. Obiekty nadpróbkowanie znajdują się zatem w części lub w całej zdefiniowanej hipersferze.

4. Opracowanie algorytmu Synthetic Majority Undersampling Technique (SMUTE) wraz z zaproponowaniem jego integracji ze znanym z literatury algorytmem SOMTE w jeden algorytm, w którym zachodzi jednocześnie proces nad- oraz podpróbkowania – Combined Synthetic Oversampling and Undersampling (CSMOUTE).
5. Opracowanie algorytmu Potential Anchoring (PA), który również wykorzystuje gaussowską radialną funkcję bazową definiowaną dla pewnego zestawu tzw. punktów bazowych. Punkty te wyznaczane są za pomocą metody klastrowania. Główną cechą algorytmu PA jest dążenie do zachowania po procesie nad- lub podpróbkowania kształtów radialnej funkcji bazowej.
6. Zaproponowanie wykorzystania algorytmów CCR oraz RBO w schemacie pozwalającym na przeprowadzenie nadpróbkowania obiektów należących do etykiet klas mniejszościowych w problemie wieloklasowym. Nadpróbkowanie nie jest przeprowadzane w przypadku obiektów należących do najbardziej licznie reprezentowanej etykiety klasy.
7. Wykonanie bardzo szerokiego zestawu eksperymentów komputerowych mających na celu weryfikację postawionej hipotezy badawczej dla każdego z opracowanych algorytmów. W przeprowadzonych eksperymentach wykorzystano publicznie dostępne benchmarkowe zbiory danych.
8. Wykorzystanie zaproponowanej metody nadpróbkowania (CCR) problemu wieloklasowego w systemie rozpoznawania raka prostaty. Zbiór danych oraz aplikacja do klasyfikacji powstała w ramach współpracy z Zakładem Opieki Zdrowotnej Diagnostyka Consilio Sp. z o.o.

Recenzowana praca ma charakter koncepcyjno-eksperymentalny. Autor zaproponował rozwiązanie problemu badawczego dotyczącego modyfikacji liczby obiektów niezbalansowanych zbiorów danych. Uzyskane przez Autora rezultaty potwierdzają postawioną na wstępnie pracy tezę badawczą, która została udowodniona w sposób eksperymentalny. Kierunek badań naukowych omawiany w dysertacji jest niewątpliwie ważny zarówno z poznawczego jak i praktycznego punktu widzenia. Wyniki o charakterze utylitarnym znajdują odzwierciedlenie w systemie do diagnostyki raka prostaty.

Uwagi krytyczne i dyskusje

Algorytm RBO przedstawiony na stronie 20 rozprawy jako jeden z parametrów wykorzystuje sąsiedztwo zdefiniowane przez *k* najbliższych sąsiadów obiektu przynależnego do etykiety klasy mniejszościowej. W ten sposób określone sąsiedztwo pozwala na wyznaczenie wartości funkcji potencjałowej. W pracach: Koziarski, Michał, Bartosz Krawczyk, and Michał Woźniak. "*Radial-based approach to imbalanced data oversampling.*" International Conference on Hybrid Artificial Intelligence Systems. Springer, Cham, 2017 oraz Koziarski, Michał, Bartosz Krawczyk, and Michał Woźniak. "*Radial-based oversampling for noisy imbalanced data classification.*" Neurocomputing 343 (2019): 19-33, które w treści dysertacji są przytoczone jako prace przestawiające rozszerzone wyniki eksperymentów związane z metodą RBO, wspomniany parametr *k* najbliższych sąsiadów nie jest parametrem algorytmu RBO.

Algorytmy RBO oraz RBU wykorzystują gaussowską radialną funkcję bazową. W dysertacji brakuje rozszerzonego uzasadnienia zastosowania radialnej funkcji bazowej typu gaussowskiego, w kontekście występowania innych typów tej funkcji oraz lokalnych charakterystyk danych.

Algorytm CSMOUTE stanowi integrację algorytmu SOMTE oraz propozycji podpróbkowania Autora dysertacji wykorzystującej pewną cechę algorytmu SMOTE. Tą cechą jest generowanie syntetycznych obiektów leżących na odcinku między wybranym obiektem należącym do etykiety klasy mniejszościowej a innym losowo wybranym obiektem należącym do tej samej etykiety klasy oraz należącym do sąsiedztwa zdefiniowanego przez *k* najbliższych sąsiadów. Cecha ta jest wskazywana w Rozdziale 2 jako wada algorytmu SMOTE, a zaproponowany algorytm CSMOUTE nie jest jej pozbawiony.

W dysertacji brakuje dyskusji dotyczącej stosowalności zaproponowanych algorytmów dla różnych skali pomiarowych, w szczególności dla cech jakościowych.

Podsumowanie

Reasumując stwierdzam, iż mgr inż. Michał Koziarski posiada ogólną wiedzę teoretyczną w zakresie metod uczenia maszynowego, które mieszczą się w dyscyplinie Informatyka Techniczna i Telekomunikacja. W szczególności posiadana wiedza dotyczy metod przetwarzania wstępnego zbiorów niezbalansowanych, algorytmów klasyfikacji, w tym konwolucyjnych sieci neuronowych oraz przeprowadzania eksperymentu naukowego. Lektura

dysertacji pozwala jednoznacznie stwierdzić, że Autor zaprezentował na jej łamach umiejętność samodzielnego prowadzenia pracy naukowej.

Postawiona na wstępie pracy hipoteza badawcza została udowodniona poprzez wykonanie wielu eksperymentów i porównaniu zaproponowanych algorytmów przetwarzania wstępnego danych niezbalansowanych z metodami referencyjnymi. Opracowane przez mgra inż. Michała Koziarskiego algorytmy stanowią oryginalne rozwiązanie problemu naukowego jakiem jest przetwarzanie wstępne danych niezbalansowanych.

Wobec powyższego, recenzowana praca spełnia wymagania zdefiniowane przez artykuł 13 ust. 1 Ustawy z dnia 14 marca 2003 r. o stopniach naukowych i tytule naukowym oraz o stopniach i tytule w zakresie sztuki (Dz. U. nr 65 r., poz. 595 z późn. zm.). Konkludując, wnoszę o przyjęcie rozprawy oraz dopuszczenie mgra inż. Michała Koziarskiego do publicznej obrony.

Dodatkowo biorąc pod uwagę dorobek publikacyjny mgra inż. Michała Koziarskiego, na który składają się między innymi publikacje w tak renomowanych czasopismach jak *Pattern Recognition*, *Knowledge-Based Systems*, *IEEE Transactions on Neural Networks and Learning Systems* czy też *Neurocomputing*, wdrożenie wyników badań naukowych w systemie diagnostyki raka prostaty, który powstał przy współpracy z Zakładem Opieki Zdrowotnej Diagnostyka Consilio Sp. z o.o. wnoszę wniosek o wyróżnienie rozprawy doktorskiej mgra inż. Michała Koziarskiego.

R. Burduk