Bydgoszcz, 17.08.2021

Dr hab. inż. Michał Choraś, profesor uczelni
Wydział Telekomunikacji, Informatyki i Elektrotechniki
UTP, Bydgoszcz

Recenzja rozprawy doktorskiej

Imbalanced data preprocessing techniques utilizing local data characteristics,

której Autorem jest Pan

mgr inż. Michał Koziarski

realizowanej na Wydziale Informatyki, Elektroniki i Telekomunikacji AGH

1. Wprowadzenie.

Niniejsza recenzja rozprawy doktorskiej, której Autorem jest Pan mgr inż. Michał Koziarski, została wykonana na zlecenie Rady Dyscypliny Informatyka Techniczna i Telekomunikacja Akademii Górniczo Hutniczej w Krakowie (Uchwała nr RD.ITiT.WIEiT.510-3/19/205/2021 z dnia 15 czerwca 2021 r.) oraz na podstawie zawiadomienia o wyznaczeniu na Recenzenta w postępowaniu o nadanie stopnia doktora podpisanego przez Przewodniczącego Rady Dyscypliny Informatyka Techniczna i Telekomunikacja, które podpisał Pan Profesor dr hab. inż. Marek Kisiel-Dorohinicki.

Promotorem niniejszej rozprawy jest Pan profesor dr hab. inż. Bogusław Cyganek, a na promotora pomocniczego powołano dra inż. Bartosza Krawczyka.

Praca jest napisana w języku angielskim.

Praca otrzymała wsparcie, m.in. z:

- Narodowego Centrum Nauki, program PRELUDIUM (grant numer 2017/27/N/ST6/01705),
- Narodowego Centrum Nauki, program OPUS (grant numer 2015/19/B/ST6/01597).

Rozprawę odebrałem 21 czerwca 2021 r., a recenzję wysłałem w wyznaczonym terminie w sierpniu 2021 r.

Niniejsza recenzja (poza wprowadzeniem i wnioskiem) zawiera odpowiedzi na siedem pytań dotyczących rozprawy doktorskiej.

2. **Jaki jest problem naukowy (teza) rozprawy? Czy został on trafnie i jasno sformułowany? Jaki charakter ma rozprawa?**

Rozprawa, której Autorem jest Pan mgr inż. Michał Koziarski dotyczy uczenia maszynowego. W szczególności, Autor zajął się rozwiązaniem problemu niezbalansowania danych. W pracy zaproponował szereg własnych metod, a jego motywacją było zaproponowanie metod otrzymujących lepsze rezultaty niż standardowe podejście SMOTE (j. ang. *Synthetic Minority Oversampling Technique*).

Autor zaproponował sześć własnych algorytmów dla podejścia binarnego, oraz dwa algorytmy dla podejścia wieloklasowego. Autor dokonał także ewaluacji zaproponowanych metod w zastosowaniu do analizy obrazów histopatologicznych.

Niniejsza praca naukowa ma charakter koncepcyjno-eksperymentalny.

Problemy naukowe rozprawy zostały jasno i trafnie sformułowane, a także rozwiązane przez Autora. Czytelna i dobrze sformułowana teza pracy znajduje się we wstępie pracy na stronie 4 (Rozdział 1.3) . Teza została potwierdzona i dowiedziona przez Autora rozprawy.

3. **Czy w rozprawie przeprowadzono w sposób właściwy analizę źródeł, w tym literatury światowej, stanu wiedzy i zastosowań w przemyśle? Czy wnioski z przeglądu źródeł sformułowano w sposób jasny i przekonywujący?**

Przegląd i analiza literatury są najsłabszą częścią rozprawy – w samej bibliografii Autor wymienił zaledwie 72 pozycje, z czego ponad 10 prac własnych. Części prac dotyczących analizy stanu zawarł w swoich artykułach związanych z rozprawą, tym niemniej w samej rozprawie mógł odświeżyć i rozszerzyć analizę stanu wiedzy.

Mimo tego faktu Autor słusznie umotywował swoje prace. Doktorant dobrze rozumie i uwzględnia aktualny stan wiedzy oraz literatury światowej w zakresie rozprawy. Autor jest bez wątpienia ekspertem w dziedzinie uczenia maszynowego i prac nad balansowaniem danych (uczeniem maszynowym na danych niezbalansowanych). Autor pracuje z naukowcami zajmującymi się tą tematyką od dawna.

Część analizowanych źródeł została wykorzystana w krótkim Rozdziale 2 rozprawy (*Preliminaries*).

4. **Czy autor rozwiązał postawione zagadnienia? Czy użył do tego właściwych metod dowodząc, że posiadł umiejętności związane z metodyką i metodologią prowadzenia badań naukowych? Czy przyjęte założenia są uzasadnione?**

Generalnie, Autor w sposób odpowiedni rozwiązał problemy, których dotyczy rozprawa. Nie mam wątpliwości, iż Autor posiada dużą wiedzę dot. zagadnień związanych z uczeniem maszynowym i balansowaniem danych. Przyjęte założenia są uzasadnione, dobrze umotywowane i merytorycznie poprawne. Teza rozprawy została dowiedziona.

Autor posiada duże umiejętności w wykorzystywaniu metod uczenia maszynowego, proponowania metod balansowania danych, implementacji swoich propozycji oraz ich testowaniu w wybranych zastosowaniach.

Autorskim, oryginalnym przyczynkiem naukowym i głównym elementem rozprawy są propozycje wielu własnych algorytmów balansowania danych, ich implementacja oraz weryfikacja oraz analiza wyników.

5. **Na czym polega oryginalność rozprawy, co stanowi samodzielny i oryginalny dorobek autora, jaka jest pozycja rozprawy w stosunku do stanu wiedzy czy poziomu nauki reprezentowanych przez literaturę światową?**

Główną częścią rozprawy są Rozdziały 3 oraz 4, zatytułowane odpowiednio:

- *Binary Resampling Strategies* (Rozdział 3),
- *Multiclass Resampling Strategies* (Rozdział 4),

w których Autor przedstawił propozycję własnych metod balansowania danych.

Autor zaproponował (i opublikował w dobrych czasopismach) aż sześć metod/algorytmów dla strategii binarnych oraz dwie metody dla strategii wieloklasowych:

- Binary resampling strategies
  - *CCR: Combined Cleaning and Resampling*
  - *RBO: Radial-Based Oversampling*
  - *RBU: Radial-Based Undersampling*
  - *RB-CCR: Radial-Based Combined Cleaning and Resampling*
  - *CSMOUTE: Combined Synthetic Oversampling and Undersampling*
  - *PA: Potential Anchoring*
- Multiclass resampling strategies
  - *MC-RBO: Multiclass Radial-Based Oversampling*
  - *MC-CCR: Multiclass Combined Cleaning and Resampling*

Wszystkie zaproponowane metody zostały odpowiednio przedstawione i omówione (z wykorzystaniem materiałów z własnych publikacji).

Postawione we wstępie pracy problemy badawcze zostały rozwiązane, a skuteczność propozycji Autora udowodniona w przeprowadzonych eksperymentach przedstawionych w publikacjach oraz w Rozdziale 5 (*Applications in Histopathology*). Autor przebadał wpływ własnych algorytmów na analizę obrazów medycznych (histopatologicznych). Eksperymenty są poprawne merytorycznie i dobrze zaprojektowane.

Warto zauważyć, iż zaproponowane przez Autora metody zostały opublikowane w czasopismach naukowych ze współczynnikiem wpływu – Autor wykazał aż 12 takich publikacji jako część niniejszej Rozprawy.

Bardzo podobał mi się fragment omawiający zależności między 12 pracami Autora umieszczony na stronach 2 i 3, pokazujący dojrzałość Doktoranta w prowadzeniu badań naukowych.

Uważam, że Autor ma bardzo dobry i bogaty dorobek publikacyjny na tym etapie rozwoju naukowego.

6. **Czy autor wykazał umiejętność poprawnego i przekonywującego przedstawienia uzyskanych przez siebie wyników? Jaka jest poprawność redakcyjna rozprawy?**

Niniejsza rozprawa stanowi przykład profesjonalnie przygotowanej pracy doktorskiej. Praca napisana jest na wysokim poziomie edycyjnym oraz graficznym. Praca jest konkretna i kompaktowa, czyta się ją bardzo dobrze.

Poziom językowy jest dobry, a w rozprawie znaleźć można nieliczne, wymagające korekty usterki gramatyczne i leksykalne, typowo dotyczące braku lub błędnego używania angielskich rodzajników (*articles*) takich jak *a, an, the*.

Te drobne usterki nie zmieniają ogólnej opinii o dobrym poziomie językowym i edycyjnym rozprawy.

Zarówno przeprowadzone eksperymenty, jak i uzyskane wyniki przedstawione są w sposób jasny i klarowny oraz metodologicznie poprawny.

Struktura pracy nie jest typowa, ale w tym przypadku odpowiednia i przemyślana. W Rozdziałach 3 i 4 wszystkie proponowane metody są opisane wg tego samego szablonu, a Rozdział 5 zawiera wyniki eksperymentów.

7. **Jakie są słabe strony rozprawy i jej główne wady?**

Rolą recenzenta jest zauważenie ewentualnych niedociągnięć i mankamentów przedstawianej pracy, oraz zgłoszenie uwag, które mogą być pomocne i przydatne w dalszych pracach.

Uwagi krytyczne, których nie mam wiele, dotyczą między innymi:

- Krótka analiza literaturowa oraz niewykorzystanie bibliografii do potwierdzenia (nawet słusznych) tez i twierdzeń.
- Umieszczanie w Rozprawie ogólnych stwierdzeń i zdań, bez ich potwierdzenia w literaturze tematu, np.:
    - *Despite their widespreadness, SMOTE-based techniques tend to be susceptible to various data difficulty factors, such as disjoint class distributions, small junctions, presence of noise, insufficient amount of training data, etc. This, to some extent, is caused by the fact that the original SMOTE algorithm does not utilize the information about majority class observations: placement of the synthetic observations generated by*

SMOTE is based solely on the relative position the neighboring minority class observations

- *The simplest data-level approaches are the unguided data preprocessing techniques, that is random oversampling (ROS) and random undersampling (RUS). However, despite their simplicity and computational efficiency, they tend to be outperformed by the guided resampling algorithms*

Zdania i twierdzenia są prawdziwe, ale jednak powinny być wzmocnione bibliografią, wynikami, poprzednimi pracami i artykułami.

- Autor nie podał procentowego szacowanego udziału w wielo-autorskich pracach będących częścią Rozprawy. Tym niemniej, poza jedną pracą, zawsze był pierwszym Autorem.
- W Rozdziale wstępnym brakuje schematu/rysunku (widoku „z góry"), który pozwoliłby na umiejscowienie prac Autora w szerszym kontekście sztucznej inteligencji, uczenia maszynowego lub analizy/klasyfikacji danych, w tym danych niezbalansowanych. Autor nie zawarł także rysunku/tabeli pozwalającego na umieszczenie jego prac w odniesieniu do innych prac w tej tematyce.
- Miejscami Autor wspomina o pewnych technologiach, metodach lub zbiorach danych, bez słowa wprowadzenia, wyjaśnienia lub nawet bez rozwinięcia skrótu. Przykładami są: pojawiające się nagle PCA oraz t-SNE na stronie 50, lub zamiana VGG16 na VGG 19 bez słowa komentarza.
- Brakuje informacji inżynierskich i informatycznych, czyli informacji o wykorzystanych technologiach informatycznych; w pracy powinien znaleźć się opis stosu technologicznego, użytych technologii itp.
- Bardzo trudno ocenić jest aktualny poziom TRL (j. ang. *Technology Readiness Level*) rezultatów pracy Autora.
- Nie jest jasne dlaczego Autor wybrał jedno zastosowanie do ewaluacji metod (obrazy medyczne)? Dość skrótowo (str. 43) Autor uzasadnił potrzebę korzystania z zaawansowanych metod balansowania danych w analizie obrazów medycznych, ale na pewno takich zastosowań/dziedzin jest więcej. Brakuje informacji o potencjalnych zastosowaniach w innych dziedzinach.

Tym niemniej, praca jest bardzo wartościowa i ma **szereg mocnych stron**, a wymienione uwagi krytyczne nie wpływają na pozytywną ocenę niniejszej rozprawy.

Autor zajął się ciekawym i bardzo potrzebnym zagadnieniem balansowania danych dla uczenia maszynowego i zaproponował aż osiem autorskich metod zarówno dla zadań klasyfikacji binarnej, jak i wieloklasowej .

8. Jaka jest przydatność rozprawy dla nauk technicznych?

Praca dotyczy bardzo aktualnych i potrzebnych zagadnień nowoczesnej informatyki technicznej i telekomunikacji. Każda praca z zakresu sztucznej inteligencji,

uczenia maszynowego, a także balansowania danych ma potencjalną przydatność w naukach technicznych i w gospodarce (w wielu możliwych dziedzinach).

Autor rozprawy zaproponował przykładowe wykorzystanie swoich metod w analizie obrazów medycznych i dowiódł ich skuteczności.

Tym niemniej, zaproponowane algorytmy mogą znaleźć zastosowanie w wielu innych dziedzinach, w tym np. w cyberbezpieczeństwie (analizie ruchu i wykrywaniu ataków sieciowych).

9. Wniosek

Biorąc pod uwagę przedstawioną przez Doktoranta rozprawę stwierdzam, że spełnia **ona wymagania stawiane rozprawom doktorskim** przez obowiązującą Ustawę i wnioskuję o **dopuszczenie** jej do publicznej obrony. Biorąc pod uwagę dorobek publikacyjny, a także wysoką jakość rozprawy, wnioskuję także o dyskusję nad jej wyróżnieniem.