Prof. Michał Woźniak
Politechnika Wrocławska
michal.wozniak@pwr.edu.pl

Wrocław, 29 listopada 2021r.

Recenzja rozprawy doktorskiej

Mateusz Lango

zatytułowanej:

Analysis of data difficulty factors for multi-class imbalanced problems and their application in classification methods

1. Problem badawczy i jego znaczenie

W pracy rozważane są problemy klasyfikacji danych niezbalansowanych, ze szczególnym uwzględnieniem trudności danych dla problemów wieloklasowych. Praca ma charakter naukowy - doktorant proponuje w pracy nowe algorytmy klasyfikacji tego typu danych oraz dokonuje ich ewaluacji na drodze eksperymentu komputerowego. Opracowane metody mają zastosowanie praktyczne, m.in. zostały zastosowane w problemach analizy wydźwięku emocjonalnego (sentymentu) wypowiedzi, ale mogą znaleźć zastosowanie w innych problemach z zakresu diagnostyki medycznej, bankowości, czy ogólnie pojętego bezpieczeństwa.

2. Wkład autora

Na początku rozprawy autor sprecyzował zakres tematyczny związany z opracowaniem metod oceny trudności wieloklasowych danych niezbalansowanych oraz ich wykorzystanie do sformułowania metod ich klasyfikacji. Doprowadziło to do sformułowania następującej tezy pracy (przytaczam jej polskie tłumaczenie zawarte w streszczeniu):
Można zaproponować nowe metody konstrukcji klasyfikatorów dla wieloklasowych danych niezbalansowanych, które będą uwzględniać informacje o źródłach trudności związanych z rozkładem danych zarówno na poziomie lokalnym jak i bardziej globalnym.
O ile cel pracy oraz zawartość rozprawy oceniam bardzo dobrze, to zwrócić należy uwagę, że teza pracy jest błędnie postawiona, gdyż jakikolwiek algorytm umożliwiający klasyfikacje spełni ten warunek. W tezie należałoby wskazać jaką przewagę (i nad jakimi algorytmami) mają uzyskać zaproponowane metody.

Do najważniejszych osiągnięć rozprawy zaliczam:

- Opracowanie algorytmu wstępnego przetwarzania niezbalansowanych danych wieloklasowych SOUP (*Similarity Oversampling and Undersampling Preprocessing*), który na podstawie, zaproponowanego przez doktoranta, wskaźnika *safe_level* usuwa z grupy obiektów należących do klas większościowych obiekty o najmniejszym poziomie wspomnianego wskaźnika oraz replikuje obiekty z klasy mniejszościowej o największej wartości *safe_level*.
- Opracowanie modyfikacji algorytmu *Roughly Balanced Bagging* dla problemu wieloklasowego – *Multiclass Roughly Balanced Bagging* (MRBBag).
- Opracowanie algorytmu *SOUP-Bagging* wykorzystującego w uczeniu klasyfikatorów bazowych perturbowane zbiory danych dodatkowo modyfikowane przez algorytm SOUP.

- Oszacowanie jakości zaproponowanych algorytmów na drodze eksperymentu komputerowego, zarówno na danych benchmarkowych, jak i na danych dedykowanych problemowi analizy sentymentu.
- Oszacowanie, dla wybranych algorytmów, złożoności obliczeniowej, co jest niewątpliwie istotne z punktu widzenia ich zastosowań praktycznych.
- Opracowanie otwartej biblioteki programistycznej w języku Python, która zawiera dwie z opracowanych metod (SOUP-Bagging i MRRBB), podczas, gdy kolejna z nich (SOUP) jest dostępna jako implementacja w języku Java.

Wyniki uzyskiwane w trakcie pracy nad rozprawą były szeroko publikowane. Doktorant jest współautorem ponad 20 publikacji, w tym wiele z nich znalazło się w materiałach liczących się konferencji międzynarodowych (w tym m.in. na ECML) i czasopismach (pięć publikacji), w większości indeksowanych w JCR. Na tym etapie kariery naukowej należy uznać dorobek doktoranta za ponadprzeciętny.

3. Poprawność

Lektura rozprawy prowadzi do sformułowania uwag o naturze ogólnej i szczegółowej.
Do uwag natury ogólnej zaliczam:

1. W rozdziale 2.2. przedstawiono kilka metryk stosowanych w klasyfikacji binarnej, w tym dla problemów niezbalansowanych, a także dwie metryki, które można stosować dla przypadków wieloklasowych. W mojej opinii warto to odnieść się do pracy Branco P. et al. (*Relevance-based Evaluation Metrics for Multi-class Imbalanced Domains*), gdzie zawarto szeroki przegląd możliwych do stosowania metryk. Należałoby tu przeprowadzić dyskusję na temat możliwości ich stosowania, a wyniki tej dyskusji wykorzystać np. na etapie badań eksperymentalnych. Doktorant nie analizuje np. wykorzystania *micro-* i *macroaveraging*, nie charakteryzuje wpływy wybranych metryk na potencjalne wnioski. Warto tu sięgnąć do prac zespołu, z którego wywodzi się doktorant, który opublikował ciekawe analizy w tym zakresie, dając m.in., wskazówki co do doboru parametrów metryk parametrycznych, m.in. w stosowanej w rozprawie średniej harmonicznej. Dobrze też jest zapoznać się z krytycznymi uwagami, m.in. Davida Handa w zakresie np. powszechnego stosowania miary F1.
2. Zawarta w rozdziale próba oceny czynników trudności jest moim zdaniem dyskusyjna. O ile zgadzam się, że wymienione w rozprawie charakterystyki danych zdecydowanie wpływają na trudność w klasyfikacji, to już próba oceny na drodze badań symulacyjnych, wykonanych przy znajomości charakterystyk rozkładów prawdopodobieństw dla danego zadania, jest obarczona pewnym błędem. Autor, wskazuje na stronie 25 (pierwszy wzór – szkoda, że wzory nie są numerowane) regułę decyzyjną algorytmu bayesowskiego, która jest prawdziwa, ale tylko dla pewnego szczególnego przypadku, tj., dla „zero-jedynkowej" funkcji strat. W przypadku ogólnej postaci funkcji strat należałoby tu wskazać na regule minimalizującą tzw. ryzyko średnie, czyli wartość oczekiwaną funkcji strat. Wskazana przez doktoranta reguła jest zatem optymalna względem *accuracy*, a w pracy wykorzystuje się ją jako rozwiązanie referencyjne dla miary zbalansowanej dokładności *balanced accuracy* - BAC (która nie została wymieniona w rozdziale 2). Zatem, jeżeli doktorant chciałby odnieść się do wyników algorytmu bayesowskiego, to należy wykorzystać kryterium optymalizacji, która uwzględni dysproporcje pomiędzy klasami, czyli warto zastanowić się nad sformułowaniem takiej funkcji strat, które te dysproporcje uwzględni. Stąd też nie jest wiarygodny wskaźnik *OCRatio* (drugi wzór na s.25), gdyż dla *recall* algorytm wykorzystujący regułę decyzyjną (pierwszy wzór na s. 25), zwany jako algorytm optymalny, a de facto optymalny dla „zero-jedynkowej" funkcji strat, nie będzie

optymalny dla problemu, gdzie ważność poszczególnych klas jest różna, a co za tym idzie nie jest optymalny z punktu widzenia *recall* i BAC.

3. Analiza jakości zaproponowanych metod jest dokonywana w pracy na podstawie badań eksperymentalnych. W moje opinii, autor powinien zadbać o większą różnorodność metod referencyjnych, np. wykorzystać jako algorytmy referencyjne, te które znalazły się w przeglądzie literaturowym. Algorytm SOUP jest porównywany z dość prostymi metodami, a zakres badań eksperymentalnych dotyczy pojedynczych metryk typu: *gmean*, *average minority* (nie opisana w rozprawie) oraz F1 (tu rozumiem *macro-averaging* F1). Eksperymenty porównawcze odnoszą się już tylko do *gmean*. Także dobór algorytmów klasyfikacji jest dość ograniczony – wybrano trzy algorytmy C4.5, 3NN oraz PART, przy czym strojenie modelu pokazano jedynie dla pierwszego z nich. Autor nie przedstawił żadnych przesłanek, dlaczego akurat te klasyfikatory zostały wybrane, w tym dwa klasyfikatory regułowe. Należałoby sięgnąć po inne klasy metod klasyfikacji, typu SVM, metody bayesowskie, czy neuronowe. W przypadku metod wykorzystujących perturbowane zbiory danych do budowy zespołów klasyfikatorów, autor właściwie ogranicza się w porównaniu do pierwotnych wersji zmodyfikowanych algorytmów, a zakres modeli stosowanych przy uczeniu klasyfikatorów bazowych także ograniczono do trzech, wcześniej wymienionych, modeli
4. Nie jest dla mnie zrozumiał sposób wykorzystania testów statystycznych. Autor przedstawia wyniki testu parowego Wilcoxona (*signed rank test*), a następnie wynik testów Friedmana z post hoc Nemenyi. Jaki to ma sens? Należy zdecydować się na jedno narzędzia analizy. W tym wypadku oparcie wniosków na analizę z wykorzystaniem jednego narzędzia – w mojej opinii testy parowe z wykorzystaniem testu Wilcoxona są wystarczające. Co więcej, w literaturze można znaleźć szereg krytycznych uwag w zakresie stosowania testu NxN Friedmana do analizy jakości algorytmów.
5. W rozdziale nr 6 zamieszczono wyniki badań dla problemu analizy sentymentu. Nie do końca jest dla mnie jasne, dlaczego eksperyment został opisany osobno, gdyż doktorant nadal wykorzystał w eksperymencie powszechnie dostępne bazy, które zostały poddane odpowiedniej obróbce wstępnej. Z jednej strony faktyczne dotyczą one pewnego spójnego zagadnienia analizy sentymentu, z drugiej mogłyby być z powodzeniem elementem badań nad oceną algorytmów bazujących na perturbowanych zbiorach danych z wykorzystaniem metod doktoranta. W opisie badań nie jest jasne, dlaczego nie przedstawiono tych samych metryk, jak w rozdziale 5. Dlaczego nie dokonano analizy wyników z wykorzystaniem *signed rank test* Wilcoxona, a ograniczono się jedynie do testu NxN Friedmana. Nie jest też jasne, czy F-score wykorzystany w badaniach, to *macro-averaged* F1 prezentowany na s.10.

Uwagi szczegółowe:

1. Brak numeracji wzorów.
2. W przeglądzie brakuje mi szerszego przeglądu metod preprocessingu danych, które starają się w jakiś sposób jednocześnie modyfikować rozkład klasy większościowej i mniejszościowej, gdyż właściwie z tej klasy pochodzi propozycja algorytmu SOUP, a także pogłębionej dyskusji na temat wad metody SMOTE.
3. J48, to implementacja algorytmu C4.5 w środowisku Weka i należy stosować raczej nazwę algorytmu, a nie implementacji.
4. W pracy wykorzystywany jest klasyfikator PART, jednak brak jest jakichkolwiek referencji do opisu tej metody (Frank E., Witten IH, *Generating Accurate Rule Sets Without Global Optimization*), której implementacji można znaleźć w Weka.
5. W algorytmie 4 (s. 60) należy wskazać z jakiego rozkładu są losowane próbki (linia 5).
6. Zakres przedstawionych badań eksperymentalnych nie jest na tyle duży, aby zamieszczać część wyników jako aneks w postaci strony internetowej.

7. Brak stosowania spójnego środowiska implementacyjnego, tj. SOUP jest zaimplementowany w Java, a pozostałe metody w języku Python.

4. Wiedza kandydata

Na podstawie lektury uważam, że doktorant posiada wiedze z zakresu informatyki, w szczególności w zakresie metod klasyfikacji obiektów, metod analizy danych niezbalansowanych, a także potrafi zaplanować eksperyment komputerowy w celu oceny jakości zaproponowanych metod.
Przegląd literaturowy dotyczący zagadnień przedstawionych w rozprawie, zawarty głównie w rozdziale pierwszym i drugim pozwala stwierdzić, że doktorant posiada aktualną wiedzę z zakresu klasyfikacji danych niezbalansowanych, a także potrafi dokonać krytycznego przeglądu źródeł w celu wskazania ciekawych kierunków badań. Zawarty w dysertacji spis źródeł literaturowych, zawierających 175 pozycji, jest aktualny i w miarę kompletny.

5. Podsumowanie

Doktorant wykazał się w recenzowanej rozprawie właściwie stosowanym podejściem eksperymentalnym oraz dobrą znajomością aktualnej problematyki związanej z projektowaniem algorytmów klasyfikacji danych. Zostało to poparte dobrymi studiami literaturowymi, obejmującymi aktualne piśmiennictwo związane z problematyką rozprawy, głównie w zakresie rozważanej problematyki klasyfikacji niezbalansowanych danych. Świadczy to o dobrej wiedzy doktoranta z tego zakresu. Dla wspomnianych problemów doktorant sformułował szereg ciekawych i użytecznych metod oraz obserwacji.
Recenzowana dysertacja przedstawia rozwiązanie ważnego i oryginalnego problemu, wzbogacając naszą wiedzę dotyczącą klasyfikacji obiektów, szczególnie w zakresie problemów wieloklasowej klasyfikacji danych o różnolicznych frakcjach obiektów należących do poszczególnych klas. Zawarte w niej wyniki badań eksperymentalnych wskazują również na możliwość wykorzystania otrzymanych metod w praktyce, co zostało potwierdzone wynikami badań eksperymentalnych dla kilku problemów analizy wydźwięku wypowiedzi.
Przedstawione w punkcie 3 recenzji uwagi nie wpływają znacząco na dość pozytywne wrażenie o przedłożonej rozprawie, a ich zamieszczenie może być przydatne dla doktoranta w przypadku szerszego publikowania wyników rozprawy, bądź poszukiwania możliwości rozwoju swoich metod.

Biorąc pod uwagę opinie zaprezentowane w poprzednich punktach i wymagania zdefiniowane przez artykuł 13 Ustawy z dnia 14 marca 2003 r. o stopniach naukowych i tytule naukowym (z późniejszymi zmianami) [1] moja ocena rozprawy pod względem trzech podstawowych kryteriów jest następująca:

A. Czy rozprawa zawiera oryginalne rozwiązanie problem naukowego? (wybierz jedną opcję stawiając znak **X**)

☐	☒	☐	☐	☐
Zdecydowanie TAK	*Raczej TAK*	*Trudno powiedzieć*	*Raczej NIE*	*Zdecydowanie NIE*

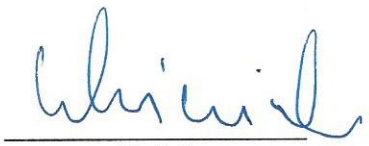
B. Czy po przeczytaniu rozprawy zgadzasz się, że kandydat posiada ogólną wiedzę teoretyczną w dyscyplinie Informatyka techniczna i telekomunikacja?

[1] http://www.nauka.gov.pl/g2/oryginal/2013_05/b26ba540a5785d48bee41aec63403b2c.pdf

[X]	[ ]	[ ]	[ ]	[ ]
Zdecydowanie TAK	*Raczej TAK*	*Trudno powiedzieć*	*Raczej NIE*	*Zdecydowanie NIE*

C. Czy kandydat posiada umiejętność samodzielnego prowadzenia pracy naukowej?

[ ]	[X]	[ ]	[ ]	[ ]
Zdecydowanie TAK	*Raczej TAK*	*Trudno powiedzieć*	*Raczej NIE*	*Zdecydowanie NIE*



Podpis