Prof. dr hab. inż. Szymon Jaroszewicz
Instytut Podstaw Informatyki PAN
ul. Jana Kazimierza 5, 01-248, Warszawa

Warszawa, 19.01.2022

Recenzja rozprawy doktorskiej

Mateusza Lango

zatytułowanej:

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

1. Problem badawczy i jego znaczenie

Rozprawa dotyczy problemu modelowania danych niezrównoważonych w przypadku obecności więcej niż dwóch klas. Przypadek ten jest istotny w praktyce a jednocześnie niedostatecznie opisany w literaturze. Problem ma bez wątpienia charakter naukowy.

2. Wkład autora

Wkład autora polega na całościowym podejściu do problemu modelowania niezrównoważonych danych w przypadku obecności więcej niż dwóch klas. Przejdę teraz do jego szczegółowego omówienia.

Autor rozpoczyna (rozdział 3) od analizy wpływu różnych charakterystyk danych uczących (takich jak nakładanie się klas) na dokładność klasyfikacji w omawianym problemie. Zadanie było znacznie bardziej złożone niż dla przypadku dwuklasowego i wymagało przeprowadzenia analizy dla różnych rodzajów nierównowagi klas, takich jak obecność wielu klas większościowych (*multi-majority)* czy mniejszościowych (*multi-minority*). Na początku rozdziału autor stawia pytania badawcze, które następnie systematycznie analizuje, dzięki czemu cały wywód został przeprowadzony systematycznie. Szczególnie interesujące są wyniki dla klas „pośrednich", które, jak pokazał autor mogą pełnić rolę zarówno klas mniejszościowych jak i większościowych. Pewną wadą rozważań autora wydaje się być dość ubogi dobór algorytmów: zastosowano jedynie drzewa decyzyjne i metodę najbliższych sąsiadów (dalsze uwagi na ten temat zamieszczę w części 3 recenzji).

W rozdziale 4 autor wykonuje kolejny logiczny krok, opracowując lokalną miarę trudności klasyfikacji danych niezrównoważonych. Autor uogólnia tu zaproponowaną wcześniej przez Promotora dla problemu dwuklasowego metodę *safe_level* wykorzystując przy tym wyniki z rozdziału 3. Na przykład, autor słusznie zauważa (str. 40), że bezpośrednie zastosowanie miary *safe_level* jest równoważne dekompozycji na problem dwuklasowy i proponuje odpowiednią korektę. Następnie autor wykorzystuje opracowaną miarę do zaproponowania algorytmu repróbkowania korygującego nierównowagę klas nazwanego SOUP. Algorytm łączy metody *over-* i *undersampling*, usuwane są tym przede wszystkim punkty zwiększające trudność klasyfikacji ustalone według zaproponowanej wcześniej miary, wykazując tym samym jej przydatność.

Autor porównuje zaproponowane metody z podejściami konkurencyjnymi na szeregu rzeczywistych zbiorów danych wykorzystując, co warto podkreślić, statystyczne testy rangowe. Proponowana metoda wypada w nich najlepiej, chociaż różnica nie zawsze jest statystycznie istotna.

W kolejnym rozdziale 5 doktorant rozszerza proponowane podejścia o metody bazujące na komitetach klasyfikatorów. Zaproponowane zostały dwa algorytmy. Pierwszy jest modyfikacją wcześniejszej metody RBBag korygującą pewne problemy z nią związane. Drugi wykorzystuje zaproponowaną przez autora metodę SOUP w kolejnych iteracjach klasycznej metody bagging. Drugi z proponowanych algorytmów jest wprawdzie dość prostą modyfikacją, stanowi jednak logiczną kontynuację wcześniejszych rozważań autora i swego rodzaju kulminację proponowanego przez Niego zestawu metod. Wypada też bardzo dobrze w eksperymentach, szczególnie na danych trudnych.

W kolejnym rozdziale autor przedstawia interesujące zastosowanie opracowanych metod do analizy sentymentu. Analiza jest przeprowadzona starannie i systematycznie. Autor zgromadził aż 12 zbiorów danych co jest rzadkością w tego typu analizach. Dla wszystkich zbiorów danych autor sporządził dwie reprezentacje: niskowymiarową opartą o słowniki i wysokowymiarową opartą o metodę word2vec (patrz uwaga w części 3). Autor, w logicznym porządku, zaczął od określenia stopnia trudności problemów, formułując przy tym kilka ciekawych wniosków, a następnie przeszedł do samej analizy. Wyniki analizy wskazują na dobre działanie modeli zaproponowanych przez autora, zwłaszcza metody SOUP-Bagging. Co ciekawe, elementami komitetu były modele liniowe, a konkretnie regresja multinomialna. Szkoda, że część 6.3 opisująca wyniki i stanowiąca swego rodzaju kulminację wysiłków autora jest krótka i dość uboga. Autor podaje link do strony zewnętrznej, zawiera ona jednak głównie 'surowe' dane i wyniki testów.

Zaproponowane w pracy rozwiązania należy z pewnością uznać za praktycznie istotne. Pokazuje to w szczególności rozdział 6, gdzie metody te zostały zastosowane do rozwiązania praktycznego problemu.

3. Poprawność

Od strony formalnej pracę oceniam wysoko. Wywody są logiczne, a stwierdzenia udowodnione eksperymentalnie. W wielu wypadkach autor stosuje systematyczne porównania algorytmów oparte o testy rangowe. Za ważny aspekt pracy uznaję też systematyczność i logiczny porządek wywodów. Na przykład, autor rozpoczyna od ogólnej analizy problemu, proponuje miarę jego trudności a następnie stosuje ją w algorytmie repróbkowania, który następnie stosuje w komitetach drzew decyzyjnych.

Praca jest przygotowana bardzo starannie i nie zawiera błędów logicznych czy nieścisłości (jeden wyjątek jest opisany poniżej). Poziom redakcyjny również oceniam wysoko. Rysunki są czytelne i dobrze opisane. Zdarzają się drobne błędy językowe, które nie wpływają jednak na czytelność rozprawy. Kilka przykładów:

str. 31: „the role ... is more sophisticated", lepiej brzmiałoby „the role ... is more nuanced"
str. 35: „manually recognized", lepiej brzmiałoby „manually determined"
str. 68: „two data representation" powinno być „two data representations"

Przejdę teraz do przedstawienia pewnych uwag krytycznych dotyczących pracy.

Moim zdaniem najsłabszą stroną pracy jest dobór algorytmów stosowanych w eksperymentach w rozdziałach 3, 4 i w porównaniach w rozdziale 5. Autor stosował głównie drzewa decyzyjne i metodę najbliższych sąsiadów, które są obecnie rzadko stosowane jako podstawowe modele predykcyjne. Na obronę doktoranta dodam, że wybór ten jest charakterystyczny dla dziedziny uczenia na danych niezrównoważonych, zapewne dlatego, że wpływ takich danych na te metody jest znaczący. Szczególnie istotne wydaje mi się nieuwzględnienie metod opartych o metodę największej wiarygodności takich jak regresja logistyczna czy sieci neuronowe z odpowiednią funkcją straty.

Jest to o tyle ważne, że zachowanie regresji regresji logistycznej na danych niezrównoważonych jest dobrze przeanalizowane teoretycznie: nierównowaga powoduje systematyczne niedoszacowanie wyrazu wolnego przy względnie dobrych wartościach pozostałych współczynników. Problemy wynikające z nierównowagi można więc skorygować prostą kalibracją modelu. Kolejnym uzasadnieniem poddania tych modeli analizie jest rozdział 6 niniejszej pracy gdzie multinomialna regresja logistyczna uzyskała bardzo dobre wyniki.

W pracy zabrakło też analiz na tzw. dużych danych, które obecnie często pojawiają się w praktyce. Chodzi tu o dane wysokowymiarowe (tysiące i więcej zmiennych) i dane o dużej liczbie rekordów. Największym zbiorem stosowanym w pracy jest zbiór *tripadvisor*, który należy uznać za niewielki według dzisiejszych standardów. W rozdziale 6 autor mówi o reprezentacji wysokowymiarowej, jednak 300 zmiennych trudno za taką uznać.

Jak pisałem, praca nie zawiera błędów logicznych czy metodologicznych. Wyjątkiem jest jednak opis wyników eksperymentów w rozdziale 5 w tabeli 5.1. Autor wybiera lepszy z algorytmów uMRBBag i oMRBBag dla każdego przypadku i tak uzyskany wynik porównuje z algorytmem RBBag*. Jest to błąd metodologiczny: wybór lepszego z algorytmów powinien być dokonywany na zbiorze walidacyjnym i dopiero taki meta-algorytm można porównywnać z RBBag*. Niemniej jednak sam algorytm uMRBBag jest wyraźnie lepszy od RBBag* więc ogólna konkluzja pozostaje poprawna.

4. Wiedza kandydata

Przegląd literatury przedstawiony w rozdziale 2 jest wystarczająco szczegółowy i dobrze ilustruje stan wiedzy w rozważanej dziedzinie. Bibliografia liczy 175 pozycji i należy ją uznać za obszerną i wyczerpującą. Nie mam zastrzeżeń do co kompletności danych bibliograficznych i ich poprawności.

Podsumowując, nie mam wątpliwości, że kandydat dobrze zna aktualny stan wiedzy dotyczącej analizy danych niezrównoważonych. Na podstawie pracy dotyczącej konkretnego tematu trudno jest ocenić *ogólną* wiedzę w dyscyplinie Informatyka techniczna i telekomunikacja (pytanie to jest wyraźnie stawiane w formularzu). Biorąc jednak pod uwagę fakt, że kandydat samodzielnie implementował proponowane metody i przeprowadził ich analizę można z dużym prawdopodobieństwem założyć, że wiedza ta jest znacząca.

5. Inne uwagi

Przejdę teraz do podsumowania. Moja ogólna ocena pracy jest wysoka. Należy przede wszystkim podkreślić, że kolejne kroki badawcze podejmowane przez autora mają dobre uzasadnienie w przeprowadzonych uprzednio analizach i stanowią logiczną kontynuację kroków poprzednich. Ułatwia to lekturę pracy i daje wrażenie jej ogólnej spójności. Praca jest napisana starannie pod względem merytorycznym i redakcyjnym.

Uzyskane wyniki są z pewnością istotne z praktycznego punktu widzenia, co pokazuje rozdział dotyczący analizy sentymentu. Praca dotyczy problemów uczenia maszynowego zbyt słabo rozpoznanych do tej pory w literaturze, wypełnia więc ważną lukę w obecnym stanie wiedzy.

Pan Lango jest autorem/współautorem ponad 20 publikacji dotyczących problemu niezrównoważenia danych i analizy sentymentu. Część z nich ukazała się w materiałach prestiżowej konferencji ECML/PKDD lub w czasopismach z *impact factor*. W przypadku wielu prac Doktorant jest samodzielnym bądź też pierwszym autorem co pozwala wnioskować, że posiada On umiejętność samodzielnego prowadzenia pracy naukowej.

6. Podsumowanie

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**)

[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?

[ ]	[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*

Jaroszewicz

Podpis

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