dr hab. Wojciech Rejchel, prof. UMK
Wydział Matematyki i Informatyki
Uniwersytet Mikołaja Kopernika

Toruń, 13.02.2026

Recenzja rozprawy doktorskiej mgr. inż. Adama Wawrzeńczyka

„Advances in positive unlabeled data modeling"

Problem klasyfikacji binarnej jest bardzo ważnym zagadnieniem w statystyce i uczeniu maszynowym. Polega on na możliwie najlepszym przyporządkowaniu obiektu do jednej z dwóch klas, bazując na pewnych cechach tego obiektu oraz dostępnych danych. W rozprawie rozważa się sytuację, gdy etykiety obiektów ze zbioru danych są tylko częściowo obserwowalne. Problem ten zwykle nazywa się modelem PU (z ang. positive unlabeled), gdyż dostępne dane dzielą się na dwa podzbiory: obserwacje etykietowane (z założenia jest to część obserwacji dodatnich) oraz nieetykietowane (pozostałe dodatnie oraz wszystkie ujemne). W oczywisty sposób analiza tego problemu jest znacznie trudniejsza niż oryginalnego zagadnienia klasyfikacji binarnej. Jednakże jest ona niezbędna, gdyż dane PU są powszechnie spotykane w medycynie, finansach, rozpoznawaniu obrazów, socjologii czy ekologii.

Omówienie rozprawy doktorskiej

Rozdział pierwszy zawiera wprowadzenie do badanej problematyki oraz krótkie przedstawienie osiągniętych wyników w rozprawie. Jej omówienie rozpocznę od rozdziału drugiego, w którym zdefiniowano model PU oraz jego podstawowe charakterystyki (na przykład częstość etykietowania czy funkcję skłonności). Opisano również problem minimalizacji ryzyka oraz ryzyka empirycznego, a także wyznaczono równoważne odpowiedniki ryzyka, które lepiej przystają do modelu PU (Twierdzenia 2.4 - 2.6).

Rodział trzeci rozpoczyna omówienie dwóch podejść do modelu PU: SS (z ang. single sample) oraz CC (z ang. case-control). W tym pierwszym zakłada się, że obserwacje są niezależnymi wektorami wylosowanymi z populacji. Natomiast w drugim dysponujemy dwoma zbiorami danych: z podpopulacji etykietowanej oraz z całej populacji. Zatem podejścia SS i CC mają różne założenia, więc bezrefleksyjne (aczkolwiek dość powszechne) użycie metod właściwych SS do CC (i vice versa) sprawia, że wyniki tych analiz są przynajmniej

częściowo bezużyteczne. W rozprawie klarownie omówiono ten problem przy upraszczającym założeniu SCAR (z ang. selected completely at random).

Zauważmy, że dane nieetykietowane w CC są mieszanką obserwacji pozytywnych i negatywnych z proporcją dodatnich $\pi = P(Y = 1)$. W rozdziale drugim wykazano, że przy założeniu SCAR dane nieetykietowane są również mieszanką tych dwóch rozkładów, ale z inną proporcją. Różnica ta jest wyjątkowo wyraźna, gdy częstość etykietowania jest duża (Rysunek 3.1). Nieadekwatność metod PU-SS do danych PU-CC (i vice versa) sformalizowano w Propozycji 3.1.

Ponadto w rozdziale tym omówiono dwa popularne klasyfikatory z podejścia CC: nieobciążony (uPU, Plessis et al., 2015) oraz jego nieujemną wersję (nnPU, Kiryo et al., 2017). Następnie używając faktów z rozdziału poprzedniego, zaproponowano nowe odpowiedniki tych metod adekwatne w schemacie SS. Na koniec przeprowadzono eksperymenty numeryczne, w których empirycznie badano tę nieodpowiedniość. W części 3.4 pochylono się nad niebanalnym problemem konstrukcji danych, które pozwolą na możliwie sprawiedliwe porównanie rozważanych metod.

Rozdział czwarty jest główną częścią rozprawy. Opisane w nim narzędzia staną się bazą algorytmów proponowanych w kolejnych rozdziałach. Rozważany jest problem ogólny (No-SCAR), gdy funkcja skłonności może zależeć od zmiennych objaśniających obiektu.

Początek rozdziału to przystępne omówienie skomplikowanych metod opartych na sieciach neuronowych, począwszy od autokoderów (AE) i wariacyjnych autokoderów (VAE), poprzez minimalizację dolnego oszacowania funkcji wiarogodności (ELBO), kończąc na generatywnym algorytmie VaDE z mieszankami gaussowskimi. Połączenie tych metod zostało użyte do danych PU w pracy Na et al., 2020 (VAE-PU). Autor znacząco poprawia ten algorytm. Najważniejsza modyfikacja (podrozdział 4.5.3) polega na użyciu wygenerowanych przez VAE-PU obserwacji pozytywnych i nieetykietowanych (P&U) do znalezienia obserwacji pozytywnych wśród dostępnych danych nieetykietowanych, podczas gdy w pracy Na et al. (2020) wygenerowane P&U były bezpośrednio użyte do estymacji ryzyka. Wspomniany zabieg pozwala uniknąć „przycinania" funkcji ryzyka, a przez to utraty informacji. Naturalnie wymaga on również dodatkowej pracy - w omawianym rozdziale została ona wykonana przez algorytmy OCC (z ang. one-class classification).

Rozdział czwarty kończą eksperymenty numeryczne, w których przewaga zaproponowanego algorytmu VAE-PU+OCC nad VAE-PU, a także innymi procedurami PU (LBE, SAR-EM) jest ewidentna.

Celem rozdziału piątego, podobnie jak części poprzedniej, jest poszukiwanie P&U w zbiorze obserwacji nieetykietowanych. Jednak tym razem badanie oparto na metodzie wielokrotnego testowania, aby uzyskany zbiór był w znacznym stopniu rozłączny ze zbiorem obiektów negatywnych. Popularne metody, na przykład procedura Benjaminiego-Hochberga kontrolująca FDR (z ang. false discovery rate), nie mogą być w tej sytuacji użyte, gdyż jedynie symulowane

obserwacje P&U, w odróżnieniu od obiektów negatywnych, są dostępne. Jest to kluczowe przy wyznaczaniu p-wartości (część 5.2.4) i zmusza Autora do oparcia analizy na kontroli FOR (z ang. false omission rate). Ten ostatni współczynnik był znacznie rzadziej badany niż FDR, dlatego na początku rozdziału piątego skoncentrowano się na rozwinięciu tego podejścia (analogicznie do kontroli FDR). Najpierw wykazano, że FOR jest asymptotycznie równoważne bayesowskiemu FOR (Twierdzenie 5.2), które jest łatwiejsze w analizie (Twierdzenie 5.3), co pozwala zaproponować klarowną regułę je kontrolującą - wzór (5.9). Metodę tę zweryfikowano w eksperymentach numerycznych, a następnie użyto do danych PU w połączeniu z procedurą VAE-PU z poprzedniego rozdziału.

Główne zadanie rozwiązywane w trzech poprzednich rozdziałach polegało na znalezieniu P&U w zbiorze obserwacji nieetykietowanych (czyli gdy $S = 0$). W naturalny sposób prowadzi to do nowego problemu badawczego przedstawionego w szóstym rozdziale: przyporządkuj nowy obiekt do jednej z dwóch klas, bazując na pewnych cechach X tego obiektu oraz stowarzyszonej z nim zmiennej S. Problem ten nazwano w rozprawie *rozszerzonym* modelem PU (z ang. augmented PU, APU). Wyniki teoretyczne dotyczące nowego modelu zawiera Twierdzenie 6.1. Pierwszy z nich mówi, że najlepszy klasyfikator w APU ma postać identyczną jak standardowy klasyfikator bayesowski, w którym prawdopodobieństwo $P(Y = 1|X)$ zamieniono na $P(Y = 1|X, S = 0)$ (przypadek $S = 1$ jest oczywisty). Interesującym rezultatem jest własność (iii), w której szacuje się różnicę ryzyka najlepszego klasyfikatora w APU i standardowego klasyfikatora bayesowskiego. Wynik ten potwiedza fakt, że użycie dodatkowej informacji (czyli zmiennej S) poprawia jakość predykcji. Co więcej, konstrukcja nowego klasyfikatora wykorzystującego tę dodatkową wiedzę nie jest wymagająca. W części numerycznej zbadano skuteczność tego algorytmu w nowym modelu APU, a także klasycznym modelu PU.

W rozdziale siódmym rozważono ciekawą i często spotykaną sytuację, gdy rozkłady danych uczących i testowych są różne. Mówiąc dokładniej, skupiono się na zadaniu, w którym rozkłady a priori klas się różnią, a warunkowe rozkłady cech w klasach pozostają niezmienne (z ang. label shift). Dodatkową trudnością rozważoną przez Autora jest częściowa obserwowalność etykiet obserwacji (model PU), a pewnym ułatwieniem jest dostępność zmiennej S w zbiorze testowym (czyli APU). Kluczowym założeniem jest (7.3), mówiące że funkcje skłonności w populacjach treningowej i testowej są równe. Głównymi wynikami teoretycznymi są Lemat 7.1 i Twierdzenie 7.4, w których pokazano, między innymi, że ilorazy szans w obu populacjach są proporcjonalne, co skutkuje faktem, że klasyfikator bayesowski w populacji testowej wymaga jedynie zmiany progu w jego odpowiedniku w populacji uczącej.

Dodatkowy problem estymacji odsetka $\tilde{\pi}$ w populacji testowej jest sprawnie rozwiązany (przy znajomości analogicznego odsetka z populacji uczącej) w (7.6). W Lemacie 7.3 oszacowano błąd estymacji tej metody i wykazano jej asymptotyczną nieobciążoność. Na koniec sprawdzono skuteczność procedury w eksperymentach numerycznych.

Uwagi

1) jak wspomniałem, poziom eksperymentów numerycznych jest wysoki. Jednak można zauważyć również pewne (nieliczne) niedociągnięcia. Jednym z nich jest mała (zwykle 10-20) liczba powtórzeń procedur na badanych zbiorach danych. Sprawia to, że średnie i odchylenia są obliczane z niewielu wartości. Domyślam się, że takie postępowanie było wymuszone przez złożoność obliczeniową rozważanych procedur. Ten ostatni problem był nieunikniony, a główny wpływ miało oparcie rdzenia wszystkich procedur na sieciach neuronowych (VAE-PU). Czy możliwe jest ich znaczące przyspieszenie (np. poprzez zrównoleglenie)?

2) zastanawiam się nad usunięciem kroku 5 w algorytmie 2, który oparty jest na wyznaczeniu p-wartości. W uproszczonej wersji używamy funkcji *score* do sortowania elementów nieetykietowanych, a następnie wybieramy odsetek p (z kroku 6) elementów z najmniejszą wartością funkcji *score*. Podejście to pozwala uniknąć dzielenia zbioru danych na zbiór treningowy i walidacyjny, co mogłoby zwiększyć skuteczność procedury,

3) użycie t-testów (chociażby tabela 4.2) jest nieuzasadnione, gdyż wymagają one danych niezależnych. Tutaj kolejne wartości są zależne, gdyż pochodzą z analizy tego samego zbioru danych,

4) zgadzam się z wymienionymi na stronie 100 potencjalnymi wadami VAE-PU+FOR, a zwłaszcza z tym że jest ona bardziej konserwatywna niż VAE-PU+OCC, co jest widoczne w Tabeli 5.5 (szczególnie w przypadku zbioru MNIST OvE). Wydaje się, że powinno to powodować drastyczny spadek jakości proponowanej procedury. Tymczasem jej skuteczność jest porównywalna z VAE-PU+OCC (Tabele 5.3 i 5.4, $\alpha = 0.1$, zbiór MNIST OvE). Jak wytłumaczyć ten fenomen?

5) niektóre wnioski wyciągane z eksperymentów numerycznych są niewłaściwe. Podam dwa przykłady:
a) Tabela 6.2: Autor sugeruje dominację VP-B+S dla małych c, podczas gdy VP+S działa bardzo podobnie. Ta usterka prawdopodobnie była spowodowana skupieniem się na mylących pogrubianych wynikach (o tym piszę poniżej),
b) w Tabelach 6.2 - 6.5 podano wyniki dotyczące skuteczności VP-B+S i jego rywali na danych symulowanych i rzeczywistych. Natomiast w umieszczonych w Dodatku D Tabelach D.1 - D.4 przedstawiono analogiczne rezultaty dotyczące zbalansowanej skuteczności. Lakoniczne podsumowanie Autora brzmi „we can see, however, that the obtained results are close to results presented in the main text". Zdanie to jest prawdziwe, gdy porównamy Tabele 6.2, 6.3 z D.1, D.2, odpowiednio. Jednakże porównanie dwóch pozostałych par tabel daje sprzeczne konkluzje. Zwłaszcza odmienne są Tabele 6.4 i D.3:
- w Tabeli 6.4 dominacja VP-B+S + true y(x) nad VP-B+S + true s(x) jest wyraźnie widoczna, podczas gdy w Tabeli D.3 skuteczność obu klasyfikatorów jest bardzo podobna dla $c = 0.5, 0.7, 0.9$,
- w Tabeli 6.4 S-Prophet wygrywa z Y-Prophet dla $c = 0.5, 0.7, 0.9$, natomiast w Tabeli D.3 sytuacja jest odwrotna,
- w Tabeli 6.4 VP-B+S wygrywa z VP-B dla $c = 0.7, 0.9$, natomiast w Tabeli

D.3 sytuacja jest odwrotna,

6) analiza przeprowadzona w rozdziale siódmym oparta jest na założeniu (7.3), które nie jest omówione w rozprawie. Wygląda ono na mocno rygorystyczne i trudne do spełnienia w ciekawych sytuacjach, to znaczy, gdy różnica między π i $\tilde{\pi}$ nie jest mała.

Pozostałe komentarze:

1) strona 21: strata l_{inv} jest rosnąca, więc nieodpowiednia w badanym problemie klasyfikacji. Prawdopodobnie miała być to strata $l(t) = 1/(1 + e^t)$ rozważana w pracy Na et al.(2020). Ponadto w Dodatku B.3 analizuje się istnienie punktów stacjonarnych ryzyka ze stratą l_{inv}, podczas gdy celem była strata z pracy Na et al.(2020). Jednakże konkluzje w obu przypadkach są takie same,

2) ryzyko empiryczne (3.10) nie jest nieobciążonym estymatorem, gdyż n_L jest losowe,

3) Algorytm 1: powinno być $R^D - R^{corr} \geq -\beta$,

4) na S w podejściu CC również można patrzeć jak na zmienną losową. Ma to konsekwencje w kolejnym punkcie,

5) strony 36-37: w podejściu SS związek między $\frac{\#\{i:S_i=1\}}{\#\{i:Y_i=1\}}$ a $c = P(S = 1|Y = 1)$ jest oczywisty. W CC ten związek jest bardziej zawiły,

6) pogrubianie wartości w tabelach miało ułatwić znalezienie najlepszej procedury w danym problemie. Jednak zostało oparte tylko na porównaniu średnich, a bez uwzględnienia odchylenia standardowego. Jest to mylące, na przykład w Tabeli 4.2 (c=0.02, MNIST 3v5) OC-SVM został wybrany najlepszym algorytmem, bo ma najwyższą średnią skuteczność. Jednakże wyprzedza IsolationForest jedynie o 0.1, a odchylenie standardowe jest większe niż 1, co podważa tę konkluzję,

7) tabela 4.7: algorytm „Baseline-modified" jest czasami znacznie wolniejszy od „Baseline-original". Dlaczego? Jeśli dobrze rozumiem, to te dwie procedury różnią się jedynie funkcją straty i sposobem łączenia obserwacji (opcja oparta na h_y zamiast h_s), co nie powinno mieć wyraźnego wpływu na czas obliczeń.

Konkluzja

Przekazaną mi do recenzji rozprawę oceniam wysoko. Uważam, że zawiera ona nowe i interesujące wyniki dotyczące analizy modelu PU w trudnym przypadku No-SCAR. Na szczególną uwagę zasługują algorytmy VAE-PU+OCC oraz VAE-PU-Bayes, które są znaczącym rozwinięciem i ulepszeniem procedury VAE-PU z pracy Na et al. (2020). Wydaje mi się, że powinny one spotkać się z dużym zainteresowaniem wśród badaczy z tego obszaru.

Ponadto chciałbym podkreślić fakt, iż eksperymenty numeryczne wywarły na mnie bardzo pozytywne wrażenie. Były dobrze przemyślane, rozległe, oparte na wielu zbiorach danych i rozważały różne ciekawe wątki badanych problemów.

Nie mam wątpliwości, że przedstawiona rozprawa doktorska spełnia ustawowe i zwyczajowe wymagania stawiane rozprawom doktorskim w dyscyplinie informatyka techniczna i telekomunikacja. Wnoszę

o dopuszczenie mgr. inż. Adama Wawrzeńczyka do dalszych etapów postępowania w sprawie nadania stopnia doktora.

Wojciech Rejchel