Prof. dr hab. inż. **Jerzy Stefanowski**
Instytut Informatyki, Politechnika Poznańska
ul. Piotrowo 2, 60-965 Poznań

Poznań, 29 stycznia 2026

Recenzja rozprawy doktorskiej mgr inż. Adama Wawrzeńczyka pt. „Advances in positive unlabeled data modeling"

1. Uwagi ogólne

Niniejsza recenzja jest sporządzona w odpowiedzi na pismo z Rady Naukowej Instytutu Podstaw Informatyki PAN. Przedmiotem oceny jest rozprawa doktorska Pana mgr inż. Adama Wawrzeńczyka **„Advances in positive unlabeled data modeling"**.

Problematyka omawiania w rozprawie mieści się w obszarze uczenia maszynowego (z j. ang. machine learning) nietypowej klasyfikacji przykładów, a dokładniej zadania tzw. uczenia się z pozytywnych i niezatykietowanych przykładów (ang. Positive – Unlabeled; w skrócie PU). Jest to dość trudne zadanie klasyfikacji, w którym, dostępna jest tylko część etykietowanych (etykieta przydziału do klasy) przykładów z klasy pozytywnej oraz zbiór przykładów nieetykietowanych (zawierający przykłady, które mogą pochodzić z klasy pozytywnej, jak i negatywnej). Jak sam Doktorant stwierdza w wstępie, że takie sformułowanie zadania uczenia klasyfikatora z ograniczoną obserwowalnością etykiet przykładów jest dość odmienne od innych, bardziej popularnych w literaturze (głównie uczenia częściowo nadzorowanego, albo uczenia się jednej klasy – ang. one class classification). Pomimo mojego własnego stwierdzenia, że jest to trochę „niszowy" problem, jednak można sprawdzić, iż prace badawcze są prowadzone, zwłaszcza w ostatnich kilkunastu latach. Doktorant zwraca uwagę, że z teoretycznego punktu widzenia ważne jest rozważanie różnych scenariuszy w jakich dostępne są przykłady z punktu widzenia rozkładów prawdopodobieństw – tzw. założenia SCAR vs. no-SCAR PU, gdyż implikuje to rozwijanie właściwych metod uczenia się. Zwłaszcza dla drugiego z nich można zauważyć w literaturze rozwój metod wykorzystujących architektury sieci neuronowych – co jest inspiracją do własnych badań Pana mgra inż. Adama Wawrzeńczyka.

W mojej opinii, taka problematyka jest ciekawa zarówno (a można nawet bardziej) z punktu widzenia badań teoretycznych, jak i prowadzi do nowych algorytmów i badań eksperymentalnych ich właściwości. Prace z tej problematyki są także publikowane ostatnio w czasopismach i na prestiżowych konferencjach. Dlatego poświęcenie jej badań i później rozprawy doktorskiej jest uzasadnione.

Wyniki prezentowane w rozprawie, na podstawie opisu zamieszczonego na końcu wstępu (patrz akapit str. 15) były publikowane przez p. A. Wawrzeńczyka i współautorów w czasopismach znajdujących się na tzw. liście Ministerialnej oraz prestiżowych konferencjach jak ECAI – co spełnia wymogi ustawowe wobec rozpraw doktorskich.

2. Uwagi na temat organizacji, stylu napisania rozdziałów.

Recenzowana rozprawa jest napisana w języku angielskim. Dostarczony maszynopis zawiera 8 rozdziałów, bibliografię, streszczenia raz dodatki z niektórymi dowodami, rozszerzoną prezentacja przykładów, eksperymentów oraz listę oznaczeń stosowanych w poszczególnych rozdziałach – łącznie 179 stron, ładnie oprawiony w stylu IPI PAN.
Konstrukcja rozprawy jest zgodna z kolejnymi osiągnięciami własnymi Pana mgra inż. Adama Wawrzeńczyka, tzn. po wstępie i dość krótkim rozdziale podstawowych pojęć oraz przeglądu

literatury nt. badan, metod dotychczas zaproponowanych dla PU, każdy kolejny rozdział przestawia osiągnięcie związane z planem opisanym zarówno w streszczeniu jak i celach rozprawy. O ile pierwsze z nich, dotyczące różnic między uczeniem się PU w różnych scenariuszach jest ciut niezależne (patrz rozdział 3), to kolejne rozdziały mają sensowny ciąg logiczny. W rozdziale 4tym wprowadzono oryginalną propozycje Doktoranta – specjalizowanego wariacyjnego autoenkodera VAE-PU+OCC – która jest podstawą do rozważania dalszych rozszerzeń zadania uczenia się z pozytywnych i niezatykietowanych przykładów PU, co Autor pokazuje w kolejnych rozdziałach, zawsze starając się najpierw rozważyć podstawy albo własności teoretyczne, później wprowadzając konkretną propozycję nowego algorytmu oraz podsumowuje ocenę eksperymentalną.

Taka konstrukcja jest dość logiczna i standardowa dla tego rodzaju badań.

Język angielski jest poprawny (nie znalazłem zbyt wielu drobnych błędów, np. formatowania), a sam styl napisania rozprawy dobry, choć w częściach teoretycznych nie tak łatwy do czytania (miejscami jest dla mnie dość hermetyczna) i wymagający pewnej wiedzy wstępnej.

Patrząc na odniesienia do literatury oraz wcześniejszych prac – odniosłem wrażenie, że Doktorant jest w nich bardzo dobrze zorientowany, czyli posiada dość rozległą wiedzę o tematyce swoich badań, oraz potrafi interesująco odnieść własne wyniki do tego kontekstu. Podsumowując, pozytywnie oceniam konstrukcje i styl napisania rozprawy.

3. Ocena wkładu oryginalnego

Rozprawa przedstawia opis pięciu osiągnięć Doktoranta.

Po pierwsze w rozdziale 3cim badano różnice pomiędzy dwoma scenariuszami dostępu do przykładów w uczeniu PU, tj. tzw single sample (gdzie dane etykietowane i nieetykietowane pochodzą z populacji o tym samym rozkładzie) i case control (gdzie to nie zachodzi). Na podstawie analizy formalnej pokazano, że zasadnicze różnice wynikają ze struktury zbioru przykładów nieetykietowanych oraz ze konieczne jest stosowanie właściwych wersji metod uczenia PU dostosowanych do odpowiedniego scenariusza (patrz np. rozdział 3.2 nt. minimalizacji ryzyka empirycznego, z którego wynika, że klasyfikator realizujący ERM dla jednego z powyższych scenariuszy nie będzie dobrze działał dla drugiego). Wprowadzono także odmianę uczenia nnPU dostosowaną do scenariusza single sample. Istotną częścią tej fazy badań są bardzo obszerne badania eksperymentalne na bogatych i zróżnicowanych zbiorach danych (chyba najpełniejsze w tej rozprawie), gdzie pokazano, że właściwe dobrane metody do scenariusza pozwoliły na wyraźną poprawę miary trafności (np. nnPUSS nad wersją nnPUCC) oraz przewaga poprawia się wraz ze wzrostem stopnia etykietyzacji w danych. W mojej opinii zawartość rozdziału pokazuje sprawność Doktoranta w analizie problemu, lecz wyniku zarówno teoretycznego jak i częściowo eksperymentu można było się spodziewać. Dla mnie dużo ciekawsze są kolejne osiągnięcia związane z modelami sieci neuronowych łączone z różnymi klasyfikatorami dla innego scenariusza dostępu do danych, tj. tzw. no-SCAR PU, gdzie rozkład etykietowanych danych niekoniecznie musi być zgodny z rozkładem przykładów klasy pozytywnej w populacji.

Za bardzo ważne i nowatorskie osiągnięcie uważam wprowadzenie metody VAE-PU+OCC – czyli podejścia generatywnego, wykorzystującego specjalizowaną wersje wariacyjnego autoenkodera i klasyfikacje jednoklasową. Rozważania dotyczą scenariusza tzw. no-SCAR problem, który jest trudniejszy niż rozważany w poprzednim rozdziale. W tym scenariuszu

Doktorant podążył drogą wcześniejszych literaturowych propozycji stosowania wariacyjnych autoenkoderów neuronowych (mogących generować nieetykietowane przykłady z elementami ich grupowania), w szczególności wcześniejszej propozycji sieci VAE-PU. Doktorant zaproponował kilka modyfikacji w rozdziale 4.5 (w funkcji straty, sposobu generowania przykładów z dopasowaniem przykładów, modyfikacji wyrażeń w empirycznym ryzyku). Natomiast za najważniejszą i najciekawszą propozycję uważam dość oryginalne połączenie z klasyfikatorem jednoklasowym (ang. One Class Classification), który realizuje pomysł, aby w zbiorze nieetykietowanym oddzielać przykłady pozytywne od negatywnych. Dodatkowo, jak sam Doktorant opisuje takie podejście eliminuje także teoretyczne problemy w podstawowym VAE-PU w formułach estymacji ryzyka. Zgodnie z moją szybką analizą propozycja VAE-PU+OCC jest nowatorska w stosunku do literatury. Ponadto doceniam jej bogatą ocenę eksperymentalną. W rozdziale 4.8 zbadano użycie różnych rozwiązań klasyfikatora OCC, i porównano z podstawowymi wersjami VAE-PU oraz dwoma metodami konkurencyjnymi z punktu widzenia różnych miar predykcji. Wyniki wskazują, że zaproponowana metoda VAE-PU+OCC przewyższa poprzednie metody (wykonano tutaj testy statystyczne, co dodatkowo wspiera ocenę przewagi), choć nie można jednoznacznie wskazać który z klasyfikatorów realizujący część OCC jest najlepszym wyborem. Ponadto nowe rozwiązane VAE-PU+OCC może działać skutecznie dla poprzedniego scenariusza etykietowania SCAR. Przedstawiona w rozdziale 4tym propozycja jest w mojej ocenie kluczowa w tym doktoracie i jest w dużym stopniu podstawą do dalszych przedstawionych osiągnięć.

I tak kolejne osiągnięcie omawiane w rozdziale 5tym, dotyczy zagadnienia kontroli wskaźnika fałszywych pominięć (FOR) w scenariuszach wykrywania wartości odstających, odnosząc się do zastosowań, w których kluczowe znaczenie ma kontrola odsetka niewykrytych wartości odstających wśród obserwacji sklasyfikowanych jako wartości prawidłowe. Z punktu widzenia wcześniejszych badań może to być innym podejściem w uczeniu PU z wariacyjnymi autoenkoderami do oceny stopnia występowania negatywnych obserwacji pośród przykładów. Jednak w moim zrozumieniu główne osiągnięcie ma charakter teoretyczny, bardziej matematyczny nt. innego spojrzenia na kontrole FOR poprzez wprowadzenie twierdzeń dowodzące, że oryginalne sformułowanie FOR i rozważana tutaj bayesowska wersja FOR są w przybliżeniu równoważne dla dużych zbiorów danych, z gwarantowanym istnieniem i jednoznacznością optymalnych progów w określonych warunkach. Ponadto wprowadzono (jak zrozumiałem nową) regułę empiryczną inspirowaną procedurą Benjamini-Hochberga, do kontroli FOR. Jest to ciekawe rozważanie teoretyczne – część z opisanych eksperymentów pokazała, że nowa procedura może kontrolować FOR przy odpowiednich dostępnych informacjach - p-wartości. Dla mnie z punktu widzenia rozprawy ciekawsza była integracja z poprzednio rozważaną siecią VAE-PU, tutaj jako VAE-PU+FOR (rozdział 5.5), lecz przedstawione wyniki eksperymentalne nie wskazały, że dla miar predykcji (accuracy i F1) uzyskano lepsze wyniki niż poprzednio wprowadzana metoda VAE-PU-OCC (w eksperymentach realizowana z Isolation Forests). Dlatego, mam trochę mieszane odczucie, co do przydatności tej metody w kontekście właściwej problematyki rozprawy.

Dla mnie dużo ciekawsze są osiągnięcia zawarte w dwóch kolejnych rozdziałach, gdzie jak zrozumiałem rozszerzono sformułowanie samego zadania PU uczenia się z ograniczoną etykietowalnością przykładów.

Pierwsze z przedstawionych rozszerzeń dotyczy nowego zadania, gdzie informacja o statusie etykietowania niektórych pozytywnych przykładów może być z czasem dostępna. Dla takiej wersji zadania Doktorant wprowadził nową regułę Bayesowską wykrywania przykładów

pozytywnych wśród nieetykietowanych przykładów. Oprócz przedstawienia jej właściwości teoretycznych zastosowano ją do rozważanego poprzednio wariacyjnego głębokiego autoenkodera – tworzą nową metodę VAE-PU-Bayes. Ciekawe są wyniki wielu badań eksperymentalnych, gdzie nowa metoda prowadzi do wyższych trafności niż porównywane wcześniejsze metody – dlatego należy docenić skuteczność nowej propozycji.

Równie ciekawe i oryginalne jest ostatnie przedstawione osiągnięcie, gdzie rozważano wariant zadania, w którym rozkłady prawdopodobieństw klas różnią się pomiędzy zbiorem uczącym a testowym. Jest to realna praktycznie sytuacja w przypadku zmiennych w czasie danych. Podobnie jak poprzednio, Doktorant przeprowadził najpierw analizę teoretyczną pokazując, że dla takiej wersji zadania należy odpowiednio zmodyfikować próg w Bayesowskiej regule decyzyjnej. W implementacji zaproponowano trzy wersje praktycznego klasyfikatora VAE-PU-Bayes / z różnymi estymacjami prawdopodobieństw, które działają dobrze w eksperymentach, które jednak teraz są skupione na wyłącznie porównaniu różnych wersji tej samej metody.

Należy podkreślić, że Doktorant udostępnił linki do repozytorium na github zawierające open source kody implementacji jego algorytmów – co bardzo doceniam.

Podsumowując moją ocenę recenzowana rozprawa zawiera wiele ciekawych propozycji rozwiązania zarówno podstawowego zadania uczenia się z pozytywnych i niezatykietowanych przykładów, które w moim zrozumieniu rozszerzają obecny stan wiedzy oraz wprowadzają nowe metody. W maszynopisie widać zarówno aspekty sprawności formalnej w analizie teoretycznej podstaw danego podzadania, poszukiwanie pewnych gwarancji teoretycznych oraz dobrego wykorzystania inspiracji z teorii, lecz również propozycje konkretnych algorytmów oraz w wielu rozdziałach ich obszerną i ciekawą analizę eksperymentalną. Ponadto jest w miarę spójna tematycznie oraz w wielu rozdziałach jest też logiczny rozwój metodologiczny przedstawianych propozycji. Dlatego oceniam ją bardzo dobrze i uważam, że z nadmiarem spełnia typowe wymagania wobec oryginalnego rozwiązania problemów naukowych w rozprawach doktorskich.

4. Uwagi dyskusyjne i polemiczne

Nie kwestionując wartości wyników zawartych w rozprawie zgłaszam poniżej kilka uwag krytycznych i dyskusyjnych. Część dotyczy wskazania tych aspektów, które można było potraktować szerzej w rozprawie lub mogą być przedmiotem dalszych badań i mam nadzieję, że otworzą dyskusję naukową podczas samej obrony, to jest:

W niektórych z opisywanych eksperymentów, można było przedstawić silniejsze motywacje decyzji, co do wyboru danych, metod i sposobów oceny. I tak w rozdziale 3cium mamy dość szeroką zróżnicowaną kolekcję zbiorów danych (patrz np. Tabela 3.1) a w badaniach różnych wersji wariacyjnych autoenkoderów w miarę spójnie użyto już tylko 4 zbiorów danych (w większości wywodzących się z tzw. benchmarków obrazowych) przetworzonych do 6 zbiorów. O ile ich opis w załączniku E wyjaśnia ich przekształcenia, aby otrzymać wektory cech, to decyzje, aby wybrać akurat te zbiory, a nie inne – np. z powyższej listy z rozdziału 3ciego nie przedstawiono – przydałoby się dodać uzasadnienie tego wyboru.

Podobną uwagę można sformułować nt. wyboru tzw. konkurencyjnych metod do porównania się. Pomimo przedstawienia różnych metod w przeglądzie w eksperymentach porównawczych, oprócz podstawowej wersji VAE-PU pojawiają się co najwyższej dwie metody LBE i SAR-EM. Znowu powtórzyłbym uwagę, że taki wybór mógłby mieć silniejszy komentarz uzasadniający

zwłaszcza, że wiele eksperymentów od rozdziału 4 do 8 pokazuje ocenę w zasadzie pewnej wersji rozważanego autoenkodera z różnymi elementami składowymi (są rodzajem tzw. ablation study tej samej metody, a nie szerokim porównaniem się z alternatywnymi rozwiązaniami – patrz np. rozdział 7, gdzie w moim zrozumieniu w zasadzie porównuje się rożne wersji estymacji w nowej regule Bayesowskiej i jej parametrów a nie ma odniesienia do żadnej alternatywnej metody). Wprawdzie odnalazłem jedno zdanie na str 67 mówiące, że z pracy (Na et al. 2020) wynika, że podstawowy VAE-PU ma być lepszy niż wiele alternatywnych wcześniejszych metod – ale chyba nie mamy gwarancji ze eksperyment wykonywano na tych samych danych i z tym samym scenariuszem oceny. Z chęcią podczas obrony poznałbym trochę pełniejsze uzasadnienie podjętych decyzji.

Powiązana uwaga dotyczy stosowanych miar oceny – w wielu eksperymentach stosuje się oprócz trafność (ang. accuracy) także miarę F1 – co trochę mnie zaskoczyło, bo dane takie jak opisane nie są raczej mocno niezbalansowane. Stąd pytanie jaka jest siła interpretacji wyniku F1, jak wydaje mi się, że on jest i tak w miarę zgodny z wynikami dla „accuracy".

Ponadto dla rozszerzonych problemów PU (rozdziały 6 i 7) większość eksperymentów jest dokonana dla miar U – obliczonych na symulowanych przykładach nieetykietowanych (domyśliłem się, ze testowych) - dla pewności dopytałbym czy w rozważanych scenariuszach podziału zbiorów – etykietowane przykładu występują także wśród testowych czy nie. Wczytując się w opis na str 112 zinterpretowałem, że celowo chciano oceniać predykcje dla nieetykietowanych – ale może pogubiłem się w interpretacji sposobów oceny, gdyż wydawało mi się, ze obecności etykietowanych przykładów klasy pozytywnej spodziewałbym się w części przykładów uczących.

Ocena kosztów obliczeń badanych metod jest przedstawiona tylko raz dla VAE-PU-OCC (Tabel 4.7, str 78). Czy dalsze z rozszerzeń tej metody mają porównywalne koszty - głównie czasy, niezależne od ew. parametrów danych, lub innej charakterystyki problemu?

Drobna obserwacja – w niektórych z badanych metod pojawiają się czasami tzw. hiperparametry, np. beta i gamma w Algorytmie 1 str 36 – w opisie eksperymentów nie udało mi się odnaleźć uwagi na temat właściwego dostrojenia ich wartości. Podobnie jak poza eksperymentalnym sprawdzeniem różnych wartości, można dobrać właściwe alpha w analizie FOR – np. dla VAE-PU+FOR w rozdziale 5 str 97 (czy wartość 0.1 jest tutaj najlepsza na podstawie przeprowadzonych eksperymentów)?

To się jeszcze wiąże z moją inną obserwacją, iż w części rozważań teoretycznych występuje „pi" proporcja klasy pozytywnej, które powinna być znana (występuje np. w oszacowaniach reguł Bayesowskich). O ile dla danych uczących mogę sobie to wyobrażać, o tyle dla nowych danych (a nie symulowanych testowych) w przypadku tylko częściowego zaetykietowania sądzę, że będzie to nieznana wartość. Stąd pytanie na ile proponowane podejścia mogłyby być odporne na niedokładne jej oszacowanie.

Z ogólniejszych pytań – czytając rozdział 4ty wprowadzający podejścia generatywne wykorzystujące wariacyjne autoenkodery (co w praktyce zdominowało dalsze propozycje Doktoranta) zastanawiałem się, dlaczego w rozważanym zadaniu PU nie pojawiały się w literaturze inne podejścia wykorzystujące inne sieci generatywne? Czy są już takie prace innych autorów, czy raczej jest to ew. kierunek przyszłych badań?

Ponadto w ramach ogólniejszego pytanie nt. dalszych kierunków badan interesowałaby mnie opinia, czy rozważane podejścia mogłyby być stosowalne, w uczeniu przyrostowym dla

zagadnienia tzw. concept drift (dryfu pojęć rozważanego w uczeniu się ze strumieni danych). Doktorat już w rozdziale 7dmy badał zmianę rozkładów klas pomiędzy zbiorem uczącym a testowym. Z racji moich własnych zainteresowań, mam skłonność do zastanawiania się na uogólnieniem przyrostowym metod bardziej statycznych, gdy dostępne dane z których uczymy klasyfikator są dostępne częściowo wraz z płynącym czasem i nadchodzące nowe dane (także uczące) mogą reprezentować zmieniającą się definicje klas pozytywnej i negatywnej (i jest to zmiana prawdopodobieństwa warunkowego $P(y|x)$, a nie tylko $p(y)$, która była rozważana).

5. Konkluzja końcowa

Recenzowana rozprawa charakteryzuje się bardzo dobrym poziomem merytorycznym i zawiera interesujące rezultaty, w tym wiele teoretycznych, które mogą być przydatne dla badań nad specyficznymi zadaniami uczenia klasyfikatorów z ograniczoną obserwowalnością etykiet przykładów. Rozważane zadanie nie jest dość typowe i łatwe, a przez to poznawczo interesujące. W zakresie rozważanych konkretnych problemów badawczych Pan mgr inż. Adam Wawrzeńczyk rozwiązał wiele zróżnicowanych, lecz powiązanych problemów badawczych (więcej niż na ogół oczekujemy wobec rozpraw doktorskich).

W mojej opinii doktorant wykazał się umiejętnościami prowadzenia badań naukowych – zwłaszcza w zakresie wprowadzania nowych metod oraz intensywnej jej oceny eksperymentalnej, pomysłowością oraz niewątpliwie pracowitością. Sam maszynopis jest interesujący dla mnie jako czytelnika i mogłem poznać wiele nowych własności rozważanych metod, które są dość oryginalne i nowatorskie. Sam maszynopis dość dobrze napisany i odróżnia się pozytywnie do wielu rozpraw, które recenzowałem.

Podsumowując, uważam, że opiniowana rozprawa Pana mgr inż. Adama Wawrzeńczyka **spełnia z nadmiarem warunki stawiane przez ustawę o tytule naukowym i stopniach naukowych w odniesieniu do rozpraw doktorskich i może być dopuszczona do publicznej obrony**.

Jerzy Stefanowski

Prof. Jerzy Stefanowski