Recenzja rozprawy doktorskiej Pana mgra Adama Wawrzeńczyka

Prof. Marek Klonowski

24 grudnia 2025

Rozprawę doktorską mgra Adama Wawrzeńczyka „Advances in Positive Unlabeled Data Modeling" **oceniam jednoznacznie pozytywnie** i wnoszę o dopuszczenie kandydata do dalszych etapów postępowania o nadanie stopnia doktora w dyscyplinie *informatyka techniczna i telekomunikacja*. Jednocześnie **wnoszę o wyróżnienie rozprawy**.

Uwagi wstępne

Praca jest zatytułowana „Advances in Positive Unlabeled Data Modeling" i ma być podstawą nadania stopnia naukowego w dyscyplinie *informatyka techniczna i telekomunikacja* w ramach nauk technicznych. Recenzja niniejsza powstała na podstawie pisma z dnia 23 października 2025, a sama rozprawa jest oceniana na gruncie ustawy z dnia 20 lipca 2018 r. – Prawo o szkolnictwie wyższym i nauce.

Tematyka rozprawy

Praca prezentuje liczne wyniki dotyczące modelu danych, które zawierają dwa typy elementów: pozytywne i negatywne – gdzie oetykietowana jest jedynie część pozytywnych elementów. Model ten jest określany jako *Positive Unlabeled* (PU). Mamy zatem sytuację taką, że część elementów pozytywnych jest ujawniona, a część jest wymieszana z elementami negatywnymi.

Taki model danych doskonale oddaje naturalne i bardzo liczne rzeczywiste scenariusze, w których uzyskujemy wiedzę o pewnych przypadkach danego zjawiska, choć nie o wszystkich. Jest to jeden z przykładów dostępu do danych, gdzie wiedza jest niepełna. Wspomniane scenariusze można *ad hoc* ilustrować przede wszystkim przykładami medycznymi. W rozprawie znajdziemy jednak szerokie spektrum innych przykładów, w szczególności dotyczących badań ankietowych.

Ze względu na powszechność zastosowań model danych PU jest szeroko badany w ostatnich latach, zarówno w ramach analiz typowo statystycznych, jak również w kontekście uczenia maszynowego. Rozpraw w mojej ocenie, nieco

wbrew deklaracjom Autora, nie dotyczy wyłącznie uczenia w modelu PU, ale stanowi ogólniejsze spojrzenie na dane tego typu.

Zawartość rozprawy

Praca składa się z siedmiu rozdziałów oraz licznych części dodatkowych. Najistotniejsze są wyniki zaprezentowane w rozdziałach 3–7, gdzie znajdziemy nowe rezultaty naukowe opublikowane pierwotnie w czasopismach i na konferencjach.

Rozprawa jest bardzo dobrze skonstruowana i logicznie uporządkowana. Szczególnie warte podkreślenia jest to, że zasadnicze rezultaty stanowią wyniki zaprezentowane w kilku powiązanych tematycznie, lecz jednak istotnie różnych pracach, a mimo to praca jest spójna. Autor nie oszczędzał swojego czasu, aby przygotować dzieło solidne, samodzielne i w pewnym sensie kompletne.

Zdecydowana większość wyników dotyczy tzw. modelu no-SCAR, w którym przyjmuje się założenie, że dane oetykietowane jako pozytywne **nie muszą** być losowo i jednostajnie wybranymi elementami pozytywnymi. Jest to założenie w oczywisty sposób naturalne dla wielu rzeczywistych scenariuszy (np. dlatego bo chorzy zazwyczaj częściej się diagnozują), które jednak krytycznie komplikuje model i bardzo ogranicza budowę użytecznych modeli teoretycznych.

Poniżej omówimy pokrótce zawartość poszczególnych rozdziałów.

Rozdział 1 przedstawia podstawowe pojęcia, anonsuje zawartość pracy, a także pokazuje powiązania treści rozprawy z poszczególnymi pracami, których współautorem jest Pan Wawrzeńczyk. Ważnym elementem jest przedstawienie problemu PU w kontekście innych, podobnych modeli badanych niezależnie.

Rozdział 2 wprowadza podstawowe formalizmy oraz solidne wprowadzenie w tematykę danych PU. W szczególności pokazuje scenariusze akwizycji danych (CC oraz SS), a także ich podstawowe własności.

Rozdział 3 Rozdział ten zawiera oryginalne, opublikowane wyniki w *Fundamenta Informaticae* i jako jedyny skupia się na modelu SCAR (*Selected Completely At Random*). Autor rozważa dwa scenariusze akwizycji danych: SS (single sample) oraz model case-control (CC). Wykazane zostaje – zarówno teoretycznie, jak i eksperymentalnie – że założenia te są istotne, a modele istotnie różne. Co więcej, ich zamienne traktowanie może prowadzić do znaczących błędów. Wskazane zostaje również, że błędy takie można znaleźć w pracach innych autorów.

Pokazane zostaje także, że różnica pomiędzy modelami silnie zależy od parametru c (frakcji oetykietowanych przypadków pozytywnych). Choć fakt, że parametr ten jest istotny, jest dość intuicyjny, należy podkreślić, że zaprezentowano tu szczegółową analizę oraz przeprowadzono bardzo liczne eksperymenty (aż 18 różnorodnych zbiorów danych).

Rozdział 4 , podobnie jak kolejne, dotyczy już znacznie atrakcyjniejszego, ale bardziej kłopotliwego, modelu no-SCAR, gdzie etykieta może zależeć nietrywialnie od cech próbki. Główną zawartością rozdziału jest rozszerzenie popularnego paradygmatu rozpoznawania klas obiektów za pomocą wariacyjnych autoenkoderów (VAE). Nowy pomysł polega na użyciu klasyfikatorów jednoklasowych (OCC) i wydaje się oryginalny na tle powiązanej literatury. Rozdział zawiera bardzo szerokie (miejscami chyba nieco nadmiarowe) opisy różnych metod oraz częściową analizę teoretyczną metody VAE+OCC.

Także w tym rozdziale widoczna jest solidna liczba zróżnicowanych eksperymentów, dobitnie potwierdzających przydatność nowego podejścia. Podczas analizy wzięto pod uwagę i zaprezentowano szczegółowo nawet czas obliczeń, nie pozostawiając wątpliwości, że zaprezentowane metody są praktyczne. Co jednak najistotniejsze, nowa metoda działa często lepiej niż wcześniejsze. Choć nie jest to regułą, w przypadkach niektórych zbiorów danych można mówić o spektakularnej poprawie.

Rozdział jest ogólnie dobrze napisany, choć niektóre fragmenty pozostawiają pewne niedosyty. Nie jest w pełni jasny pseudokod, w szczególności sposób oceny zbieżności algorytmu. Wyniki tu zaprezentowane zostały opublikowane na ECAI 2023.

Rozdział 5 prezentuje wyniki z ICCS 2023, które dotyczą metod wykrywania obserwacji odstających (ODD) za pomocą podejścia FOR (*false omission rate*). Chociaż praca silnie wykorzystuje pomysły innych autorów (szczególnie podejście FDR), wydaje się, że mgr Wawrzeńczyk musiał również w tej części wykazać się znaczną pomysłowością. Autor przedstawia wykorzystanie opracowanych metod FOR w standardowym VAE-PU, przy czym podstawowa idea może mieć znacznie szersze zastosowania.

Zrozumienie działania metod w tym rozdziale jest trudniejsze dla czytelnika, co po części wynika z częstych odwołań do innych prac, jednak całość prezentacji jest ciągle dobra. Bardzo pomagają liczne przykłady oraz dobry opis eksperymentów wykonanych na bardzo zróżnicowanych danych.

Rozdział 6 Wprowadza nowy model rozszerzonego PU (*augmented PU*), w którym udostępniana jest etykieta podczas predykcji. Przedstawiona została reguła bayesowska oraz pokazano, jak nowe podejście wykorzystać wraz z VAE-PU. Rozdział ten wymagał większej sprawności rachunkowej. Jak poprzednie, zawiera obszerne eksperymenty numeryczne na zróżnicowanych danych. Model oraz eksperymenty są bardzo dobrze umotywowane, choć nieco niejasny pozostaje wybór eksperymentów w punkcie 6.4. Wyniki z tego rozdziału zostały opublikowane na ECAI 2024.

Rozdział 7 Wprowadza – prawdopodobnie wcześniej niebadany – model label shift dla danych PU. Choć analiza ogranicza się do rozszerzonego modelu PU, wkład jest znaczący. Przedstawiono liczne wyniki eksperymentalne zarówno na

klasycznych zbiorach danych, jak i na danych syntetycznych. Wyniki z tego rozdziału zostały opublikowane w czasopiśmie AMCS.

Części dodatkowe Ostatnia część rozprawy zawiera bardzo zgrabne podsumowanie rozprawy wraz z opisem potencjalnych dalszych ścieżek badawczych. Opis ten wydaje się niewymuszony i dobrze przemyślany, a postawione problemy są dobrze umotywowane i koncepcyjnie nietrywialne.

Dalej znajduje się lista symboli użytych w rozprawie, która jest bardzo pomocna i potwierdza szeroki zakres przedstawionych wyników. Podobnie można powiedzieć o obszernej bibliografii obejmującej zarówno prace klasyczne, jak i najnowsze rezultaty. W rozprawie znajdziemy także apendyks; szczególnie dobre wrażenie robi część A.2, gdzie autor przedstawia dość żmudne rachunki w sposób niezwykle szczegółowy.

Ocena rozprawy

Ocena wyników rozprawy, jak również jej formy, jest **jednoznacznie pozytywna**.

Przede wszystkim należy zauważyć, że praca zawiera bardzo liczne i znaczące wyniki. Prezentacja wyników jest w większości bardzo dobra. Przedstawione są często nawet elementarne przekształcenia. Niektóre fragmenty są trudniejsze do zrozumienia, ale i tu widać duży wysiłek, aby ułatwić zrozumienie czytelnikowi prezentowanych idei.

Z pełnym przekonaniem stwierdzam, że wkład merytoryczny zawarty w rozprawie z dużym nadmiarem spełnia wymagania stawiane nawet bardzo dobrym rozprawom doktorskim.

Trzeba też zaznaczyć, że wyniki zostały dobrze opublikowane. Na szczególną uwagę zwracają dwie prace z ECAI (napisane wraz z promotorem). Jest to konferencja bardzo dobra i selektywna. *Fundamenta Informaticae* też w mojej ocenie gwarantuje rzetelny poziom recenzencki.

Nie mam wątpliwości, że uzyskane rezultaty są bardzo wartościowe. Badania są świetnie umotywowane naturalnymi, praktycznymi scenariuszami, a część wyników wymagała znacznej sprawności rachunkowej oraz biegłego posługiwania się aparatem matematycznym.

Doskonałe wrażenie robi również część eksperymentalna. Wyniki przedstawiono na bardzo szerokich i różnorodnych zbiorach danych, obejmujących różne typy danych (wizualne, tekstowe, tabelaryczne). Na uwagę zasługuje także udostępnienie i solidne udokumentowanie kodów oraz deklaracja powtarzalności wyników.

Uzasadnienie wniosku o wyróżnienie

Główną przesłanką wniosku o wyróżnienie jest po prostu to, że oceniana rozprawa wyraźnie wyróżnia się na tle typowych dysertacji z informatyki – i to na plus.

Praca jest bardzo solidna metodologicznie, zarówno od strony matematyczno-formalnej, jak i eksperymentalno-obliczeniowej.

W pracy znajdziemy bardzo liczne wyniki będące kontynuacją istotnych dla praktyki i teorii badań, a także nowe, znaczące pomysły. Zaprezentowane wyniki znacznie wykraczają poza typowe dla doktoratu rozwiązania postawionego problemu naukowego. W tym przypadku można mówić o widocznym wkładzie w dyscyplinę (nawet jeśli żaden z wyników/koncepcji nie jest przełomowy).

Zauważone pomyłki

Poniżej zebrano drobne, mało znaczące uchybienia redakcyjne; lista ta z pewnością nie jest kompletna, bo mimo zachwytów nad rozprawą, drobnych usterek dostrzegłem sporo.

- Na s. 32, równania 3.3 oraz 3.4 – brakuje relacji równości.
- W wielu miejscach w rozdziale 4 brakuje znaków interpunkcyjnych.
- Nie zawsze konsekwentne wprowadzanie symboli (np. $\perp$, symbol "#"jako liczność zbioru, określenie „super uniform"). Część z nich znajdziemy na liście symboli.
- Brak konsekwencji w zapisie wartości oczekiwanej oraz funkcji charakterystycznej $\mathbf{I}$.
- Bernoulli (s. 83) – prawdopodobnie powinno być binomial.
- Na s. 87: „alpha" $\rightarrow \alpha$.
- Niektóre symbole wprowadzane są później niż ich pierwsze użycie.
- Przykłady zapisane kursywą bez wyraźnego uzasadnienia (np. str. 84).
- Str. 105 (6.5), powinno być $\Delta(d)$?
- Literówki obecne już od „Streszczenia".