Kraków, dnia 24 czerwca 2021 roku

Dr hab. inż. Tomasz Hachaj
Uniwersytet Pedagogiczny im. Komisji Edukacji Narodowej
Wydział Nauk Ścisłych i Przyrodniczych
Instytut Informatyki
ul. Podchorążych 2, 30-084 Kraków

# Recenzja Rozprawy Doktorskiej mgr inż. Jędrzeja Biedrzyckiego pt. „Integration of decision trees in geometric space"

Promotor: dr hab. inż. Robert Burduk
Politechnika Wrocławska
Wydział Elektroniki
Katedra Systemów i Sieci Komputerowych

Niniejsza recenzja została opracowana w oparciu o uchwałę Rady Dyscypliny Naukowej Informatyka Techniczna i Telekomunikacja nr 51/04/DNND03/2021-2024 z dnia 21 kwietnia 2021 podpisanej przez prof. dr hab. inż. Michała Woźniaka, Przewodniczącego niniejszej Rady; która to (uchwała) wyznacza mnie na Recenzenta. Otrzymałem pocztą zawiadomienie z dnia 23-04-2021 informujące mnie o podjęciu powyższej uchwały w postaci załącznika nr 5b do ZW 109/2020. Zawiadomienie podsiał prof. dr hab. inż. Andrzeja Ożyhar, prorektor ds. Nauki Politechniki Wrocławskiej.

## 1. Ocena wyboru tematy i tez rozprawy

W pracy postawiono i starano się wykazać prawdziwość następującej hipotezy badawczej:

*„Wykorzystanie granic decyzyjnych wytrenowanych drzew decyzyjnych pozwala na zbudowanie klasyfikatora łączonego o większej wartości miary jakości klasyfikacji niż klasyfikator łączony jakim jest las losowy lub głosowanie większościowe korzystające z tego samego zbioru wytrenowanych drzew decyzyjnych."*

Autor Rozprawy doprecyzował poszczególne elementy tezy wykorzystując popularne miary jakości, których ścisłą definicję zawarł w trzecim rozdziale Dysertacji. W calu wykazania prawdziwości tezy pracy, zrealizowano kilka celów szczegółowych, których opis stanowi treść kolejnych rozdziałów Pracy.

Tematyka recenzowanej Rozprawy dobrze lokuje się w obszarze zainteresowań współczesnej informatyki, w szczególności w zakresie nadzorowanych metod uczenia maszynowego. **Wybór problemu, tezy rozprawy oraz jej celów oceniam pozytywnie.**

## 2. Ocena zawartości rozprawy

Przedstawiona do recenzji Praca składa się ze 132 stron maszynopisu. Zawiera spis treści, spis ilustracji, spis tabel, indeks skrótów wraz z wyjaśnieniami, wstęp, pięć numerowanych rozdziałów, spis pozycji literatury, który zawiera 134 pozycje, dwa załączniki (A i B) zawierające wyniki badań

przedstawione w formie tabel oraz streszczenie dysertacji w języku angielskim i polskim. Całość rozprawy napisana jest w języku angielskim, z wyjątkiem wzmiankowanego wcześniej streszczenia w języku polskim.

Układ i zawartość rozdziałów merytorycznych jest zasadniczo poprawna. We wstępie autor w kilku zdaniach zarysowuje tematykę doktoratu oraz ogólnie przedstawia podjętą problematykę badawczą. We wstępie przedstawiona zostaje również hipoteza badawcza rozprawy oraz przyczyny, dla których Doktorant zainteresował się tą tematyką. Przedstawiony jest również układ pracy.

W pierwszym rozdziale zdefiniowano problematykę klasyfikacji jako metodę uczenia maszynowego pod nadzorem. Wprowadzono szereg definicji i formalizmów, do których Autor odwołuje się w kolejnych rozdziałach pracy.

W drugi rozdziale Doktorant definiuje problematykę geometrycznego podejścia do integracji klasyfikatorów uzupełnioną o niezbędny przegląd literatury oraz przedstawia popularny algorytm AdaBoost.

W rozdziale trzecim Doktorant omawia zestaw danych, który zgromadził na potrzeby przetestowania zaproponowanych algorytmów i empirycznego udowodnienia przedstawionej tezy. Zgromadzony zestaw danych składa się dwudziestu ośmiu zróżnicowanych zbiorów danych pobranych z popularnych repozytoriów UCI oraz KEEL. Dane mają kolumny o wartościach rzeczywistych i są zróżnicowane zarówno pod względem liczby obiektów (próbek), które zawierają, liczności klas, zbalansowania (liczności obiektów, należących do poszczególnych klas) oraz liczby wymiarów. Autor przedstawia charakterystyki poszczególnych zbiorów danych w tabeli. Dzięki zastosowaniu ogólnodostępnych, darmowych zbiorów danych eksperymenty przygotowane przez autora mogą być łatwo powtórzone, co jest faktem bardzo pozytywnym. Szkoda natomiast, że autor dopiero na 36 stronie wspomina po raz pierwszy, że wszystkie eksperymenty będą przeprowadzone na dwuwymiarowych zbiorach danych. Doktorant podaje również niewiele szczegółów dotyczących uzasadnienia tej decyzji a zgoła żadnych informacji, które dwie cechy zostały wyselekcjonowane dla każdego z 28 zbiorów przedstawionych w Tabeli 3.1. Autora załącza również implementacje zaproponowanych przez siebie metod w postaci kodów źródłowych i udostępnia je na platformie GitHub. Jest to godne pochwały działanie, które może przyczynić się do popularyzacji zaproponowanych algorytmów w społeczności osób zainteresowanych takimi zagadnieniami. Szkoda natomiast, że żadne z repozytoriów nie posiada opisu – w obecnej chwili w praktyce jedynie czytelnicy Dysertacji mają szansę do nich trafić i skorzystać z ich zawartości. Rozdział trzeci zawiera również opis metodologii porównawczej oceny skuteczności klasyfikatorów.

Rozdział czwarty zawiera szczegółowe opisy zaproponowanych algorytmów, wyniki porównawcze ich działania oraz dyskusję wniosków. Duża część tego rozdziału pracy oraz otrzymanych wyników oparta jest na wcześniejszych publikacjach Autora, które znajdują się w sekcji bibliograficznej na pozycjach [13][14][15] oraz [16]. Podczas lektury rozdziały czwartego można odnieść wrażenie, że Autor mógł dołożyć większych starań, aby tak zredagować tę część pracy, aby unikać niepotrzebnych powtórzeń, szczególnie tych dotyczących selekcji cech oraz wstępnego przetwarzania danych, które jest bardzo zbliżone dla każdego z Algorytmów 3-6. O ile takie osobne wprowadzenia mogłyby być konieczne w osobnych publikacjach, to w jednolitym tekście Dysertacji mogą wprowadzać pewne nieporozumienia. Szczegółowo na ten temat piszę w czwartej części tej recenzji (Uwagi dyskusyjne i krytyczne).

Ostatni merytoryczny rozdział jest podsumowaniem całości Rozprawy. Autor uzasadnia tu spełnienie zakładanej hipotezy badawczej i podsumowuje zarówno badania jak i autorskie algorytmy.

## 3. Oryginalne osiągnięcia autora rozprawy

W recenzowanej pracy zawarto kilka wartościowych i oryginalnych koncepcji oraz rozwiązań, dokonano ich implementacji oraz uzyskano wyniki wzbogacające naszą wiedzę w zakresie omawianych zagadnień. Do najważniejszych osiągnięć autora należy zaliczyć:

1. Opracowanie, analizę, weryfikację oraz implantację przedstawionych w czwartym rozdziale algorytmów.
2. Rozbudowaną analizę statystyczną o charakterze porównawczym zaproponowanych metod w oparciu o szeroki zestaw danych.
3. Dbałość o powtarzalność przeprowadzonych badań poprzez udostępnienie implementacji w repozytorium GitHub oraz wykorzystaniu powszechnie dostępnych zbiorów danych.

Wymienione osiągnięcia są oryginalne i znaczące. Na nich zatem przede wszystkim opieram ogólnie pozytywną ocenę rozprawy. Udowodniają one postawioną tezę rozprawy oraz stanowią wkład w rozwój algorytmów nadzorowanego uczenia maszynowego dla zbiorów danych o cechach należących do dziedziny liczb rzeczywistych.

## 4. Uwagi dyskusyjne i krytyczne

Jak już wcześniej wspomniałem, przedłożona do recenzji rozprawa pod względem merytorycznym i redakcyjnym napisana jest w większości poprawnie. W mojej ocenie Autor nie ustrzegła się jednak pewnych braków i nieścisłości. Moje uwagi dyskusyjne oraz krytyczne będę wymieniał w kolejności chronologicznej, zgodnie porządkiem wystąpienia uciążliwości oraz nieścisłości w Rozprawie.

1. Pierwszą uciążliwością, która od razu rzuca się w oczy jest brak numeracji wszystkich wzorów oraz definicji, które występują w pracy. Wzorów nie jest przesadnie dużo i nie ma żadnych merytorycznych ani redakcyjnych przesłanek, aby każdy formalizm miał swój unikalny numer, do którego można by się odwołać. Brak takich unikalnych odwołań zmusza do stosowania formy opisowej w postaci numeru strony oraz przedmiotu definicji czy równania, do czego będę musiał się uciec pisząc tę recenzję. Autor ponumerował tylko te formalizmy, do których ponownie odnosił się w tekście.
2. Na stronie 15 Autor opisując problematykę klasyfikacji użył pojęcie podobieństwa (similarity) nie definiując, czym faktycznie jest to podobieństwo. Dopiero na stronie 18 pojawia się informacja o odległości pomiędzy obiektami i wymienione są z nazwy przykładowe metryki Euklidesa, Manhattan oraz Czebyszewa. Przy wcześniejszej wiedzy, że problematyka doktoratu będzie dotyczyć klasyfikacji obiektów w przestrzeni liczb rzeczywistych ten wybór metryk może wydawać się intuicyjny, natomiast czytelnik dowiaduje się o tym stosunkowo późno. W tym kontekście zdanie: „each class contains similar objects, whereas objects of distinct classes are dissimilar" jest bardzo nieprecyzyjne i bez formalnej definicji, na czym polega to "podobieństwo" można je interpretować na różne sposoby, również wbrew intencjom autora.
3. Również na dole strony 15 autor, moim zdaniem w niezbyt szczęśliwy sposób definicje zestaw funkcji $\{g_1, g_2, \ldots, g_K\}$, które przypisują prawdopodobieństwo do każdej z etykiet klasy nie dbając o to, aby sumaryczne prawdopodobieństwo było równo 1.

4. W rozdziale 1.1.1 Autor wymienia metody selekcji cech i redukcji wymiarowości. Skoro tematyka Rozprawy dotyczy uczenia pod nadzorem powinna tam również zostać wspomniana liniowa analiza dyskryminacyjna (linear discriminant analysis, LDA).
5. Strona 17 – wszystkie definicje, w których występuje logarytm powinny jawnie wykluczać sytuację, w której zlogarytmowana zostanie liczba mniejsza lub równa zero.
6. Strona 18 – nie wytłumaczono, co oznacz $E_{before}$ oraz $E_{after}$.
7. Strona 18 - "it is essential to find a smaller set of prototypes" – mniejszy, ale od czego?
8. Strona 18 – „noisy objects" – w tym kontekście chodzi zapewne o „outliers".
9. Strona 21 – Autor zakłada, że funkcja aktywacji zwraca wartości w przedziale [0,1] a następnie jako jej przykłady przedstawia funkcje ReLu oraz Leaky ReLu, które nie spełniają tych założeń.
10. Strona 22 – błąd językowy: „convolutional neural networks – composed of one or more convolutional layers, uses"
11. Strona 22 – "recurrent neural networks (...) powerful in natural language processing" – moim zdaniem, jeżeli Doktorant nigdzie później nie używa sieci rekurencyjnych ani nie odwołuje się do literatury, która o nich traktuje nie warto o nich wspominać, szczególnie w tak mało ścisły sposób.
12. Strona 26 – „Without loss of performance in classification, the process of training a model can be sped up to 20 times" – z kontekstu nie jest jasne, względem czego następuje to dwudziestokrotne przyspieszenie.
13. Strona 35, Algorytm 2 – symbol, który został użyty do oznaczenia etykiet jest bardzo podobny do symbolu, który został użyty do oznaczenia wag. Ten fakt utrudnia analizę algorytmu.
14. Strona 36 – nie podano, jakie dokładnie metody selekcji cech (wymiarów) użyto a jedynie ogólną informację, że opis tych metod znajduje się w [50, 106]. Na stronie 62 i 71 Autor podał, że zastosował do tego ANOVA. Ponieważ przy okazji każdej z czterech metod klasyfikacji zaproponowanych przez Autora w jakiś sposób dochodzi do selekcji cech (nie jest to podane w sposób jednoznaczny i jednolity dla każdego z czterech zaproponowanych przez Doktoranta algorytmów), już na tym etapie należałoby sprecyzować sposób doboru cech. Brakuje też informacji, dlaczego autor w ogóle dokonuje wybory niektórych cech. Czy nie wszystkie cechy zawierają istotne informacje? Czy zastosowanie większej liczby cech w jakiś sposób niekorzystnie wpływa na wyniki zaproponowany w Dysertacji algorytmów? Jeżeli zmniejszenie wymiaru zbiorów danych jest koniecznością, dlaczego dokonywany jest wybór cech a nie np. redukcji wymiarowości przy pomocy PCA albo LDA? Jeżeli w każdym z czterech zaproponowanych algorytmów w identyczny sposób (to tylko moje przypuszczenie – z treści pracy to nie wynika) dochodzi do selekcji cech, właśnie w rozdziale 3 powinno znaleźć się tego merytoryczne uzasadnienie wraz z koniecznymi wyliczeniami.
15. Strona 41 – we wzorze na wielkość $dist_B$ zostało użyte oznaczenie $x_0$, które zostało użyte wcześniej na stronie 18 – czy pełni ono tą samą funkcję, co poprzednio?
16. Strona 41 – „X denotes the classification space, i.e. the cube generated by the possible linear combinations of vectors in the feature space" – zapewne chodzi tu o kombinację liniową wektorów bazy układu współrzędnych, natomiast nie jest jasne, co Autor definiuje jako przestrzeń klasyfikacji i dlaczego ma ona być sześcianem.
17. Strona 42, Rysunku 4.1.1 w pracy nie ma (błąd w tekście), jest Rysunek 4.1. Nie jest jasne, co oznaczają opisy w legendzie, zapewne niebieski to wykres funkcji gdy $\beta=0.5$ a pomarańczowy, gdy $\beta=0$. Opisie tego rysunku pojawia się również sformułowanie „mass center distance", choć do tej pory autor używał pojęcia „centroid". Czy to te same pojęcia?
18. Strona 43 – na początku sekcji 4.1.2 Autor pisze, że dokonana została selekcja cech w celu zmniejszenia liczby wymiarów do tych, które zawierają najwięcej informacji. Doktorant nie

napisał, jaka metoda została do tego użyta a co za tym idzie, co decyduje o liczbie informacji zawartej w cesze. Z dalszej części pracy można się domyśleć, że chodzi tu o wariancję. Jeżeli rzeczywiście chodzi o wariancję, należy przedstawić podstawowe wyniki analizy wariancji i sprawdzić, jaka ilość wariancji została utracona w wyniku tych operacji. Redukcja albo selekcja (albo jedno i drugie - autor używa tych pojęć wymiennie) została dokonana w celu „uniknięcia niepotrzebnego skomplikowania (albo złożoności)" – na czym miałoby polegać to „niepotrzebne skomplikowanie"? Problem ten został już poruszony w 14 uwadze krytycznej do niniejszej Dysertacji.

19. Strona 43 oraz wszystkie kolejne obliczenia, w których autor odnosi się do wyników algorytmu Random Forest – autor porównuje wyniki swoich algorytmu do klasyfikatora Random Forest, który to algorytm, niezależnie od jego implementacji, ma wiele parametrów konfiguracyjnych, takich jak np. liczba drzew, różne parametry dotyczące losowego rozkładu cech i próbek w poszczególnych drzewach, ich głębokości itp. Szczegółowe informacje o ustawieniach parametrów tego algorytmu powinny znaleźć się w Dysertacji.
20. Strona 46, rysunek 4.2 – co oznacza CD – czy to critical difference?
21. Strona 49 oraz załącznik A – Autor przeprowadza dodatkowe eksperymenty na rozszerzonym zbiorze danych, natomiast nie przeprowadza dyskusji wyników ani tak dokładnej analizy statystycznej, jak miało to miejsce w rozdziale 4.1. W związku z tym czytelnik nie wie, czy osiągnięte wyniki potwierdzają czy też nie wcześniejsze obserwacje.
22. Strona 50 – Autor pisze "In the article, a two-dimensional classification problem is considered" a następnie "The dataset (including training and testing subsets) generates a N–dimensional space" - ilu wymiarowe są rozważane zbiory danych? Na tej samej stronie znajduje się sformułowanie "Those cubical sets (further referred to as subspaces) are of the same shape as the original dataset" - co Autor rozumie pod pojęciem kształtu zbioru danych?
23. Strona 52 a następnie strona 54 ("This means that $M^N$ subspaces will be generated for the N-dimensional problem), Algorytm 4, punkt 5 (Compute the coordinates of the $M^n$ static classification regions) – Autor niekonsekwentnie oznacza liczbę wymiarów raz wielką, raz małą literą. Dodatkowo, ponieważ liczba wymiarów znajduje się w wykładniku potęgi, dla odpowiednio dużej liczby wymiarów algorytm zaczyna być bardzo obciążający pamięciowo i obliczeniowo. Dla przykładu: przy najmniejszej rozważanej przez Autora liczbie podziałów wzdłuż każdej osi (20) i dla stosunkowo niewielkiego, jeżeli chodzi o liczbę wymiarów zbioru wr (#f = 11) powstanie $20^{11}$ regionów, których współrzędne, zgodnie z punktem 5 Algorytmu 4 trzeba policzyć (i zapewne przechować). Ta sytuacja ma decydujące znaczenie o możliwości praktycznego zastosowania tego algorytmu, szczególnie obecnie, kiedy wektory cech generowane przez popularne implementacje głębokich sieci neuronowe mają kilkaset wymiarów np. Facent 128 wymiarów (Schroff et. all. "FaceNet: A Unified Embedding for Face Recognition and Clustering"), VGG-16 512 wymiarów (Simonyan et. all. "Very Deep Convolutional Networks for Large-Scale Image Recognition") itd. Doktorant na stronie 56 pisze: "The algorithm is easily applicable to any number of dimensions without any modifications (...)". W świetle zaprezentowanych wcześniej faktów „łatwość zastosowania" zaproponowanego algorytmu dla wysoko wymiarowych problemów jest co najmniej dyskusyjna. Podobnie sytuacja występuje w wypadku Algorytmu 5 oraz 6, w którym liczba obszarów jest funkcją liczby drzew (K), liczby podziałów („gałęzi") drzewa decyzyjnego oraz liczby wymiarów zbioru danych (N). Dyskusja pesymistycznej złożoność obliczeniowej i pamięciowej Algorytmów 5 i 6 mogłaby ostatecznie sprecyzować te fakty.
24. Strona 52, zaraz przed równaniem (4.4) – wielkość $f_m(R_l^k)$ nie jest nigdzie zdefiniowana.

25. Strona 56 „The problem of generalization of majority voting introduced in [16] was further developed in [13]. "– niezgodność chronologiczna, praca [13] ukazała się wcześniej, niż praca [16].
26. Strona 60 oraz załącznik B – mój komentarz jest analogiczny jak w uwadze 21: Autor przeprowadza dodatkowe eksperymenty na rozszerzonym zbiorze danych, natomiast nie przeprowadza dyskusji wyników ani tak dokładnej analizy statystycznej, jak miało to miejsce w rozdziale 4.1. W związku z tym czytelnik nie wie, czy osiągnięte wyniki potwierdzają czy też nie wcześniejsze obserwacje.
27. Strona 61, również strona 37 – Autor używa sformułowania „open-source datasets" oraz „open-source benchmarking datasets", które wydaje się połączeniem dwóch odrębnych pojęć: open (benchmarking) datasets oraz open source software.
28. Strona 63 a następnie strona 71 – pojawia się niezrozumiałe sformułowanie, że prostokątne obszary mają maksymalną powierzchnie – jak liczona jest ta powierzchnia i skąd wiadomo, że jest maksymalna?
29. Strona 71 – Doktorant pisze: „The algorithm presented in this work is an improvement of the algorithm presented in the previous work" – nie pisząc, o jakie wcześniejsze prace chodzi. Przedmiotem mojej recenzji jest Dysertacja Doktoranta a nie jego wcześniejsze publikacje.
30. Strona 71 – pojawia się pojęcie pierścieni regionów (rings of regions), które nie jest nigdzie w jasny i bezpośredni sposób zdefiniowane.
31. Niektóre pozycje bibliografii są niedopracowane i utrudniają odnalezienie źródła; są to między innymi: pozycja [60] nie ma pełnych danych (brakuje informacji o monografii: Information Technologies in Biomedicine, Volume 4. Advances in Intelligent Systems and Computing, vol 284. Springer, Cham.); [75] brakuje wydawcy oraz ISBN (wiley, ISBN: 978-0-470-90874-7); [116] brak idntyfikatora pracy, jest to zapewne arXiv:0904.3664; [123] jest zapewne zasobem internetowym https://www-users.cs.umn.edu/~kumar001/dmbook/firsted.php itd.
32. Szkoda, że autor nie przedstawia dalszych perspektyw badawczych podjętej tematyki.

Należy zaznaczyć, że wymienione powyżej uwagi nie wpływają w sposób bardzo istotny na poznawcze oraz utylitarne wartości zaproponowanych rozwiązań i metodologii ich oceny.

## 5. Podsumowanie

Przytoczone powyżej uwagi polemiczne nie umniejszają wartości merytorycznej pracy, która stanowi oryginalny wkład Autora w zagadnienia związane z konstruowaniem i trenowaniem algorytmów klasyfikacji opartych na łączeniu grup "słabych" klasyfikatorów w celu budowy pojedynczego, "silnego" klasyfikatora (ensemble learning). Autor skupił się niemal w całości nad wykorzystaniem do tego celu granic decyzyjnych wytrenowanych drzew decyzyjnych.

Podsumowując recenzję stwierdzam, że moja generalna opinia o Rozprawie doktorskiej „Integration of decision trees in geometric space" mgr inż. Jędrzeja Biedrzyckiego jest pozytywna. Uważam, że przedstawiona do recenzji Rozprawa zawiera samodzielne i oryginalne, zaproponowane przez Autora rozwiązanie trudnego i aktualnego problemu badawczego, o potencjalnie licznych praktycznych zastosowaniach. Przedstawiona teza rozprawy została dowiedziona a postawione cele i zadania pracy zrealizowane. Odpowiada to ustawowym wymaganiom stawianym rozprawom doktorskim. **Na tej podstawie wnioskuję o dopuszczenie Rozprawy do publicznej obrony w celu uzyskania przez jej Autora stopnia doktora w dziedzinie nauk inżynieryjno – technicznych w dyscyplinie naukowej informatyka techniczna i telekomunikacja.**

Tomasz Hachaj