

Wydział Matematyki i Nauk Informatycznych

Politechnika Warszawska

Warszawa, 12.11.2025 roku

dr hab. inż. Agnieszka Jastrzębska, prof. PW

Wydział Matematyki i Nauk Informatycznych

Politechnika Warszawska

ul. Koszykowa 75, 00-662 Warszawa

RECENZJA ROZPRAWY DOKTORSKIEJ

Pana mgr. Tomasza Kloneckiego

pt:

„Cost-constrained feature selection using information theory”

opracowana dla Rady Naukowej IPI PAN w dziedzinie nauk inżynieryjno-technicznych,
w dyscyplinie informatyka techniczna i telekomunikacja

Promotor rozprawy: Pan dr. hab. Paweł Teisseyre

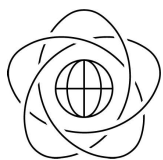
- 1. Jakie zagadnienie naukowe/badawcze jest rozpatrywane w pracy (cel i teza rozprawy) i czy zostało ono dostatecznie jasno sformułowane przez Autora?**

Rozprawa doktorska dotyczy zagadnienia selekcji cech przy uwzględnieniu podanych explicite ograniczeń budżetowych na używane cechy.

Autor w sposób precyzyjny zdefiniował cel pracy, którym było opracowanie metod umożliwiających wybór cech nie tylko pod kątem maksymalizacji dokładności predykcji, lecz także z uwzględnieniem ich kosztów. Założenia zostały sformułowane w sposób klarowny. Motywacja podjęcia tego wątku badawczego została nakreślona w sposób przekonujący – skupienie się w dużej mierze na danych pochodzących z pola medycyny ułatwia wdrożenie się w tematykę rozprawy, a równocześnie demonstruje ważną rolę podejmowanych badań. Rozwój metod integrujących aspekty selekcji cech, ich kosztu oraz dokładności rozpoznawania pozwala uzyskać rozwiązania bardziej adekwatne do realnych potrzeb analizy danych w wielu obszarach.

Autor konsekwentnie rozwija postawiony cel poprzez przedstawienie i analizę trzech uzupełniających się podejść, tworzących spójne, kompletne i dobrze uzasadnione ramy teoretyczne oraz praktyczne dla zadania selekcji cech uwzględniającej aspekt ich kosztu.

Po pierwsze, Autor opracował nową metodę selekcji cech opartą na dolnym ograniczeniu informacji wzajemnej do zastosowania w przypadku klasyfikacji jednoetykietowej, jak i wieloetykietowej. Po drugie, zaproponował metodę selekcji cech z kosztami dla grup cech. Po trzecie, przedstawił podejście do modyfikacji istniejących metod selekcji związanych



Wydział Matematyki i Nauk Informatycznych

Politechnika Warszawska

z modelami opartymi na karach (takich jak Lasso, czy MCP), tak aby uwzględniały one koszty cech w procesie uczenia.

2. Czy w rozprawie przeprowadzono w sposób właściwy analizę źródeł, w tym literatury światowej, stanu wiedzy i zastosowań w przemyśle?

Syntetyczny opis stanu wiedzy w dziedzinie, w której pracuje Autor, został zawarty w rozdziale 2. Został on podzielony na trzy części, kolejno wprowadzające w obszar selekcji cech w kontekście uczenia z nadzorem (rozdział 2.1), problematykę powiązania kosztu z cechami (rozdział 2.2) oraz klasyfikacji wieloetykietowej (rozdział 2.3). Dodatkowo rozdział 2.2 wprowadza taksonomię podejść do zadania wyboru cech uwzględniającego ich koszt dzieląc je na metody oparte o konkretny model oraz bez powiązania z modelem.

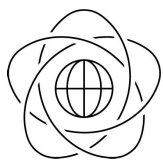
Niezwyczajnie mocnym atutem rozprawy doktorskiej są rozdziały 3 oraz 4, które w sposób precyzyjny oraz metodyczny przedstawiają matematyczne podstawy algorytmów konstruowanych przez Autora. Wywód w tych dwóch rozdziałach opiera się na u wspólnionej notacji matematycznej, wątki zostały uporządkowane w sposób logiczny. Nie ma zbędnych wtrąceń opisowych, a za to są przejrzyste przykłady. Sposób opracowania tych rozdziałów pokazuje nie tylko wiedzę, ale również umiejętność jej przekazywania przez Autora rozprawy.

3. Czy Autor rozwiązał postawione zagadnienia, czy użyto właściwej do tego metody i czy przyjęte założenia są uzasadnione?

W rozprawie omówiono nową metodę sekwencyjnej selekcji cech z ograniczeniem kosztowym. Wprowadzenie współczynnika indeksującego opartego na dolnym ograniczeniu informacji wzajemnej oraz uogólnienie popularnego kryterium Joint Mutual Information na przypadki wieloetykietowe i interakcje wyższego rzędu stanowią istotny wkład Autora. Następnie, omówiono problem selekcji cech z kosztami dla grup cech. Autor proponuje różne podejścia, z których każde ma odrębne zalety praktyczne. W końcowej części rozprawy przedstawiono modyfikację wybranych istniejących metod selekcji cech opartych o funkcję kary powiązanych z konkretnym modelem klasyfikującym w celu uwzględnienia kosztu cech.

Wymieniony zakres prac był odpowiednio szeroki. W rozprawie opracowane algorytmy są przedstawiane w sposób teoretyczny, a następnie omawiane są ich właściwości w oparciu o eksperymenty empiryczne. Eksperymenty dotyczyły zarówno danych rzeczywistych, jak i syntetycznych. Dobór metod nie budzi zastrzeżeń.

Autor naturalnie pomija część wątków funkcjonujących w obszarze algorytmów selekcji cech, jak metoda Relief, czy metody oparte o boosting gradientu, natomiast ograniczenie się do podejść związanych z teorią informacji było bardzo dobrym i racjonalnym wyborem. Nie zostały również poruszone wątki przetwarzania wstępnego danych, np. obsłużenia



Wydział Matematyki i Nauk Informatycznych

Politechnika Warszawska

obserwacji odstających, które w praktyce nie da się pominąć chcąc wyciągać wnioski z eksperymentów empirycznych. Krytyka ta nie odnosi się do warstwy teoretycznej rozprawy, a jedynie empirycznej.

4. Na czym polega oryginalność rozprawy, co stanowi samodzielny i oryginalny dorobek Autora, jaka jest pozycja rozprawy w stosunku do stanu wiedzy i poziomu techniki reprezentowanych przez literaturę światową?

Prowadzone przez Autora prace badawcze skupiały się na trzech wątkach:

(i) Opracowanie nowej metody selekcji cech, która uwzględnia informacje o kosztach poszczególnych cech. Metoda ta oparta została na funkcji oceny wyprowadzonej z wykorzystaniem dolnego ograniczenia informacji wzajemnej. Metoda jest parametryzowana, co zwiększa jej uniwersalność. Rozważany jest ogólny przypadek obejmujący klasyfikację jednoetykietową, jak i wieloetykietową. Jest to najbardziej rozbudowana część pracy.

(ii) Opracowanie nowej metody selekcji cech, która uwzględnia informacje o kosztach cech poprzez założenie, że koszt zostaje przypisany nie pojedynczej cesze, ale łącznie pewnemu zbiorowi cech.

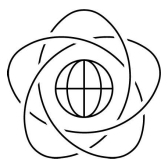
(iii) Zaproponowanie modyfikacji wybranych istniejących metod selekcji cech opartych o funkcję kary powiązanych z konkretnym modelem klasyfikującym w celu uwzględnienia kosztu cech. Takie podejście wymusza, że model jednocześnie optymalizuje dokładność predykcji oraz koszt cech. Autor skupił się na popularnych technikach, jak Lasso, Adaptive Lasso, czy Minimax Concave Penalty (MCP).

Przedstawione rozwiązania uzupełniają aktualny stan wiedzy w obszarze algorytmów selekcji cech i stanowią cenny wkład autora w tę dziedzinę. Mając na uwadze praktyczną wagę zagadnienia podniesionego w rozprawie, prowadzone prace mają potencjał dalszego rozwoju.

5. Czy Autor wykazał umiejętność poprawnego i przekonującego przedstawienia uzyskanych przez siebie wyników (zwięzłość, jasność, poprawność redakcyjna rozprawy)?

Rozprawa została przedstawiona w postaci tekstu w języku angielskim, którego ostatnia strona ma numer 116. Składa się z 8 numerowanych rozdziałów, jednego dodatku oraz spisu literatury.

Rozprawa została napisana w sposób zwięzły. Język pracy jest bardzo dobry. Wysoko oceniam układ poszczególnych jej części: prowadzony wywód został ułożony logicznie, proporcje tekstu do ilustracji i tabel odpowiednie. Chciałabym zwrócić uwagę na staranny aparat matematyczny, którym Autor wykazał się w rozprawie. Notacja jest spójna i przejrzysta.



Wydział Matematyki i Nauk Informatycznych

Politechnika Warszawska

Autor nie uniknął błędów edycyjnych, których szczegóły zostały przedstawione w kolejnym punkcie tej recenzji.

6. Jakiek sabe strony i wady rozprawy?

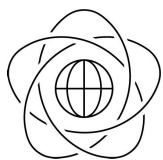
W tej czści przedstawiam uwagi szczegłowe podzielone wedł typu usterki.

Uwagi merytoryczne:

- s. 28 – „While our focus in this study is on discrete variables, it is worth noting that analogous definitions can be extended to quantitative variables.” – Zmienne dyskretne mog by ilociowe (np. liczba pasażerw w autobusie).
- s. 31 – Brak kompletnego objanienia do pierwszego wzoru.
- Posługiwanie si skrtami nazw wlasnych, ktre nie zostaly uprzednio objanione jest błędem, np. s. 17 JMI (wyjanienie jest za pzno), s. 85 BIC oraz AIC.
- s. 49 – Rys. 5.1 oraz jego opis w grnej czści s. 50. wskazuj, że dopiero w momencie przekroczenia budżetu T ma miejsce przerwanie algorytmu dodajcego cechy. Jest to niezgodne z intuicj przekazywan na pocztku pracy, np. Rys. 1.2, gdzie koszt $\leq T$.
- s. 52 – Dla kompletnoci opisu miara Hamminga powinna by zdefiniowana przed jej pierwszym użyciem w Tabeli 5.1. Należało rwnież okreli czym jest „satysfakcjonujca” miara Hamminga (zdania u dołu strony nie s jasne).
- s. 58 – Brak informacji o liczbie powtrzeń eksperymentu, ktry wizał si z losowaniem.
- s. 59 – „This Experimental Section” – Nie jest jasne do czego si to odnosi.
- s. 59 – Wyniki dla zaledwie dwch zbiorw danych rzeczywistych s doc skromnym podparciem do wycigania wnioskw.
- s. 84 – „the validation set” – Zakładam, że chodzi jednak o zbir testowy.
- s. 92 – Wyjanienia odnonie ROC AUC s o wiele za pzno.

Uwagi edycyjne dotyczce tekstu:

- W rozprawie przyjęto niespjny styl odwoływania si do wzorw, np. s. 32 kursyw „Equation (3.7)”, s. 36 sam numer „(3.5)” lub „(4.1)”, s. 37 skrt „Eq. (4.1)”.
- W rozprawie licznie wystpuj sieroty (błd edycyjny łamania tekstu), np. litera „a” na s. 15, s. 21, s. 28, s. 45, s. 46, s. 78, s. 80, s. 100, czy litera „A” na s. 24, s. 70 oraz wdowy, np. numer tabelki na s. 93.
- Błdy interpunkcyjne, a w szczegłonoci brak przecinkw (s. 29 po „Below”, s. 33 po „Below”, s. 58 po „Then”, s. 70 po „In this chapter”, s. 74 „In this section”, s. 77 po „Then”) oraz użycie minusa zamiast długiej pauzy (s. 41, s. 80).
- Małe wartoci w jzyku angielskim zwykł si pisa słownie, jak np. s. 38 („4 variables”, „3 variables”).
- s. 19 – „... in the Table at the ...” – Powinien zosta podany konkretny numer tabeli.
- s. 59 – Pierwsze zdanie powinno by rozbite na dwa.



Wydział Matematyki i Nauk Informatycznych

Politechnika Warszawska

- s. 29 – Styl pisowni tytuł podrozdziału nie jest spójny ze stylem reszty tytułów: drugi wyraz nie powinien być pisany z dużej litery.
- s. 32 – Duże litery przy „Interaction Information” są niespójne ze stylistyką pisowni pozostałych nazw w pracy.
- s. 41 – Powinno być „...the higher its value, the stronger the emphasis...”.
- s. 59 – Pisownia miar accuracy, precision, recall jest raz z dużej, a raz z małej litery.
- s. 59 – Pisownia ROC AUC obok siebie jest raz kursywą, a raz bez.
- s. 78 i in. – W tekście odwołując się do podrozdziałów czasem występuje „Section”, czasem „Chapter” – niespójność stylu.

Uwagi edycyjne do wzorów:

- Brak kropki: s. 31 pierwszy wzór, wzór (4.10), s. 43 wzory pierwszy i czwarty, wzór (6.9).
- Kropka zamiast przecinka: wzór (5.3).
- Brak przecinka: s. 48 wzór ostatni.

Uwagi do odwołań do literatury:

- Bardziej przyjaznym stylem cytowania w niektórych miejscach byłoby podanie nazwisk autorów, np. s. 23 „... the authors of [165] ...”.
- Odwołania do literatury na s. 25 powinny być połączone i posortowane (7. linijka tekstu).

Uwagi do ilustracji:

- s. 62 vs. s. 71 i in. – Podpisy obrazków są raz pod a raz nad obrazkiem.
- s. 71 – tytuł wykresu (Rys. 5.9) u góry jest zbędny mając na uwadze jego właściwy opis pod spodem.

7. Podsumowanie

Przedstawione w rozprawie konstrukcje algorytmiczne uzupełniają aktualny stan wiedzy w obszarze metod selekcji cech i stanowią cenny wkład autora w tę dziedzinę.

Reasumując stwierdzam, że rozprawa doktorska Pana mgr. Tomasza Kloneckiego, pt. „Cost-constrained feature selection using information theory” spełnia wymogi stawiane pracom doktorskim, które zostały ujęte w art. 13 ust. 1 ustawy o stopniach naukowych i tytule i naukowym oraz o stopniach i tytule w zakresie sztuki z dnia 14 marca 2003 roku (Dz. U. z 2017 r., poz. 1789) oraz art. 179 ustawy z dnia 3 lipca 2018 r. przepisy wprowadzające ustawę Prawo o szkolnictwie wyższym i nauce (Dz.U. z 2018 r. poz. 1669) i wnioskuje do Rady Naukowej IPI PAN o dopuszczenie Pana mgr. Tomasza Kloneckiego do publicznej obrony rozprawy doktorskiej.

Podpis jest prawidłowy

Digitally signed by
Agnieszka Jaszczyńska
Date: 2025.01.12
21:04:41 CET