

dr hab. inż. Ryszard Klempous prof. PWR
Wydział Informatyki i Telekomunikacji
Politechnika Wrocławska
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław

Wrocław, 05.04. 2022 r.

Tytuł **rozprawy**: Machine learning method for training survival models with applications in medicine

Autor: Wojciech Jarmulski

Opiniowana rozprawa napisana jest w j. angielskim. Liczy ona 135 stron. Rozprawa składa się z 7 rozdziałów podzielonych w mojej opinii na trzy części oraz Bibliografii, Spisu Rysunków, Spisu Tabel, Skróatów i Notacji. Zawiera streszczenia w j. polskim oraz w j. angielskim. Bibliografia rozprawy obejmuje 142 ważniejsze pozycje literaturowe należące do przedmiotu rozprawy. Pierwsza część (rozdziały 1-3) zawiera wprowadzenie do tematyki rozprawy, charakterystykę stanu wiedzy w obszarze rozprawy oraz tezy rozprawy. Część druga składająca się z rozdziałów 4-5 to przedstawienie metod użytych w rozprawie, ich weryfikacja i ocena. Oryginalne rezultaty prezentowane są w części trzeciej (6 i 7) z wykorzystaniem medycznych baz danych. Rozdział 7 zawiera uwagi końcowe (podsumowanie i wnioski) oraz kierunki dalszych prac.

Metody analizy danych medycznych były rozwijane początkowo w naukach medycznych i biologicznych. Analiza przeżycia umożliwia określenie czasu, jaki upłynie do (pierwszego) wystąpienia pewnego zdarzenia w oparciu o zbiór metod statystycznych badających zachodzące w zdarzeniu procesy. Ze względu na duży wolumen danych, koniecznością jest zaproponowanie rozbudowy metod innych niż tradycyjne metody statystyczne do tego celu. Wykorzystywanie metod uczenia maszynowego (ML) daje dobre wyniki predykcyjne, jednak o ograniczonej interpretowalności. Dlatego też pożądane jest poszukiwanie nowych metod modelowania przeżycia cechujących się pełną interpretowalnością oraz o dużej mocy predykcyjnej przeżycia. Rozprawa została zrealizowana w ramach współpracy Autora z Warszawskim Uniwersytetem Medycznym i w aspekcie aplikacyjnym dotyczyła analizy przeżywalności pacjentów po przeszczepie wątroby.

Niniejsza praca ma następujący siedmiorozdziałowy układ:

Rozdział 1 przedstawia problem oraz tezy badawcze rozprawy. **Rozdział 2** wprowadza zagadnienia i notacje obszaru analizy przeżycia, w tym popularny model proporcjonalnego ryzyka Coxa oraz miarę jakości predykcyjnej modelu przeżycia - C-index. **Rozdział 3** ma charakter przeglądu wiedzy na temat ML nakierowanego na problemy rozważane przez Autora. Dotyczy metod uczenia maszynowego, które zostały zaadoptowane do analizy przeżycia, a których wydajność jest konkurencyjna w stosunku do regresji Coxa. Autor, rozpoczynając od ogólnej wiedzy ML, przechodzi do dwóch opisów znanych metod lasów przeżycia oraz sieci neuronowych przeżycia. **Podrozdział 3.1**, mylnie nazwany przeglądem uczenia maszynowego, zawiera krótką definicję uczenia maszynowego według Friedmana oraz jego podział ze względu na typ uczenia. W rozdziale można znaleźć nieistotne wtrącenia takie jak „In the era of big data, enormous amount of data is generated every second”. Nie wszystkie wzory są numerowane. **Podrozdział 3.3** Autor rozpoczyna stwierdzeniem „In order to adopt machine learning models to survival analysis, a special version of the likelihood has to be used” bez odniesienia do literatury. Czym według autora jest *special version of the likelihood*?

W **Rozdziale 4** wyprowadzona została nowatorska metoda ML dla modeli przeżycia. Pozwala ona na osiągnięcie wysokiej mocy predykcyjnej **porównywalnej z modelami o pełnej złożoności**. przy zachowaniu pełnej interpretowalności modeli, tj. możliwości określenia wpływu każdego predyktora na wynik. Osiągnięto to modyfikując metodę wzmocnienia gradientowego w wariancie składnikowym (*component-wise gradient boosting*) i dostosowując ją do problemów analizy przeżycia poprzez optymalizację funkcji straty w postaci błędu średniokwadratowego przeżycia. Dodatkowo wykorzystane zostały modele bazowe (*base learners*) w postaci zespołu drzew (*bagged trees*) przy ograniczeniu wynikowej formy funkcyjnej modelu do sumy addytywnych funkcji nieliniowych

RK

pobierających na wejściu pojedynczy predyktor lub ich parę. Jest istotą innowacyjnego podejścia Autora.

Rozdział 5 zawiera empiryczną ewaluację komponentów zaproponowanej metody oraz jej porównanie na rzeczywistych zbiorach danych z innymi metodami uczenia maszynowego dostosowanymi do analizy przeżycia, pokazując wysoką moc predykcyjną zaproponowanej metody. Wykorzystano rzeczywiste publiczne zbiory danych: SUPPORT, WHAS, METABRIC, RGB, STD, RETINOPATHY, GBSG2, FLCHAIN. Przeprowadzono następujące eksperymenty:

- Porównanie wpływu wyboru funkcji straty na predykcję przeżycia mierzoną indeksem C na wszystkich zbiorach danych. Określono optymalne wartości parametrów modeli. (Tabela 5.1 Charakterystyka zbiorów danych)
- Porównanie wpływu funkcji straty na moc predykcyjną mierzoną indeksem C na wszystkich zbiorach danych. W celu porównania wpływu funkcji straty na moc predykcyjną, wykorzystano model gradientowego boostingu o pełnej złożoności. Stosowano różne funkcje straty.

Funkcja straty średniokwadratowego błędzie przeżycia dała najlepsze wyniki predykcyjne w większości przypadków i dlatego w proponowanej przez nas metodzie jest ona zalecaną i domyślną funkcją straty dla optymalizacji. Badano również wpływ wyboru funkcji uczącej bazowej na wydajność predykcyjną wynikowych modeli gradientowych boostingowych mierzoną indeksem C.

Rozdział 6 zawiera rezultat klasycznej medycznej analizy statystycznej przeprowadzonej na udostępnionym przez Warszawskim Uniwersytecie Medycznym (M UW) zbiorze danych pacjentów po przeszczepieniu wątroby. Zaproponowany model oparty jest na modelu proporcjonalnego ryzyka Coxa i wykorzystuje statyczną wartość AST (*transaminaza asparaginianowa*), zmienność PLT oraz trend PLT (*płytki krwi*) i WBC (*białe krwinki*). Daje on lepsze wyniki od ugruntowanego modelu MELD (*Model for End-Stage Liver Disease*) oraz modeli zbudowanych tylko z wykorzystaniem statycznych wartości pomiarów biochemicznych. Wynikająca z modelu informacja, że zmienność i trend PLT mierzonego w okresie pierwszego roku po przeszczepieniu wątroby są istotnymi predyktorami długoterminowej przeżywalności pacjentów, jest wykorzystywana w praktyce przez lekarzy.

Rozdział 7 zamyka rozprawę sumując jej zawartość, przedstawia wyniki pracy. Potwierdzono główną tezę tej pracy (**patrz zestawienie w tabeli 7.1**). **Osiągnięto wysoką moc predykcyjną przy zachowaniu pełnej interpretowalności modelu**. Skorzystano ze zbioru danych pacjentów z Kliniki Immunologii, Transplantologii i Chorób Wewnętrznych oraz pacjentów po przeszczepie wątroby (LTx) w Klinice Chirurgii Ogólnej, Transplantacyjnej i Wątroby oraz Klinice Chirurgii Ogólnej i Transplantacyjnej WUM w latach 1996-2015. Zbiór danych wejściowych zawierał informacje o 1095 pacjentach z łączną liczbą 48 772 obserwacji. Z 874 kwalifikujących się pacjentów (tj. posiadających co najmniej jedną zarejestrowaną obserwację) wykluczono tych, którzy nie spełniali kryteriów eksperymentu projektowego, tj. przeżyli minimum 1 rok po LTx i mieli dostępne dane z minimum pięciu obserwacji w pierwszym roku po LTx. Ostateczna kohorta liczyła $nn = 450$ chorych. Do tych danych zastosowano proponowaną metodę na zbiorze danych rzeczywistych pacjentów po przeszczepie wątroby. Uwzględnienie miar zmienności i trendu zwiększa dokładność modelu w przewidywaniu przeżycia pacjentów po LTx. Zaproponowano dokładny model przeżycia wykorzystujący statyczną wartość AST, miarę zmienności PLT oraz miarę trendu WBC i PLT w pierwszym roku po transplantacji jako silne predyktory długoterminowej śmiertelności.

Potwierdzono tezy pomocnicze - zaproponowano model oparty na regresji Coxa wykorzystujący statyczną wartość AST, zmienność PLT oraz trendy PLT i WBC. Przewyższa on ustalony Model for End-Stage Liver Disease (MELD). oraz modele wykorzystujące jedynie miary statyczne. Zmienność i trend pomiarów PLT w pierwszym roku po transplantacji są silnymi predyktorami długoterminowej śmiertelności. Jest już wykorzystywana przez lekarzy w ich codziennej praktyce klinicznej

P, K

W rozprawie została sformułowana teza główna i tezy pomocnicze. Dla uniknięcia wariantowości tłumaczenia tezy przedstawiam w języku rozprawy.

Main Thesis: Survival machine learning method derived by modifying component-wise gradient boosting method, optimizing the survival mean squared error loss function, using bagged trees as base learners and limiting the outcome functional form to the sum of additive non-linear functions of single predictors and their pairwise interactions, is:

- *Adjusted to survival analysis,*
- *Maximizing predictive power, producing better results than Cox regression model and is competitive with black-box full-complexity survival machine learning methods,*
- *Fully interpretable, i.e., contribution of each predictor on the outcome can be directly shown,*
- *Well suited to practical clinical applications.*

Auxiliary theses:

- *Usage of variability and trend measures of biochemical observations improves models' predictive power of survival of patients after liver transplantation.*
- *Cox model using the static value of AST, variability of PLT, and trends of PLT and WBC is better than models built only on static measures and the established Model for End-Stage Liver Disease (MELD).*

Teza formułuje metodę ML przeżycia uzyskaną poprzez modyfikację składowej metody wzmacniania gradientu, optymalizację funkcji straty średniokwadratowej przeżycia poprzez użycie predyktorów i ich interakcje w parach. Założona optymalizacja funkcji przeżycia jest realizowana poprzez uczenie modelu na danych rozpatrywanego zjawiska. Następnie użycie bagging trees do tego, aby móc korzystać z właściwości predykcyjnych drzew decyzyjnych. Największym problemem związanym z drzewami decyzyjnymi jest ich duża wariancja. **Problem, który to stwarza, polega na tym, że każda niewielka zmiana danych może spowodować poważne zmiany w modelu i prognozach na przyszłość. Jedną z zalet bagged trees jest to, że pomaga zminimalizować zmienność przy jednoczesnym utrzymywaniu stałego odchylenia. Natomiast jednym z głównych problemów związanych z bagged trees jest to, że są one niezwykle trudne do interpretacji.**

Zastosowanie straty średniej kwadratowej błędu przeżycia do zaadoptowania sieci neuronowych do problemów analizy przeżycia zostało dokonane w 2019. Metoda wykorzystania wzmocnienia gradientowego w wariancie składnikowym adoptowanego do analizy przeżycia w celu wygenerowania modeli addytywnych złożonych z sumy nieliniowych funkcji pojedynczych predyktorów i ich interakcji parami jest autorskim pomysłem Autora.

Główna teza tej pracy została potwierdzona - w porównaniu z innymi metodami uczenia maszynowego analizy przeżycia (patrz zestawienie w tabeli 7.1). Zaproponowane podejście do modelowania analizy przeżycia pozwala na osiągnięcie wysokiej mocy predykcyjnej na równi z modelami o pełnej złożoności. Zostaje zachowana pełna interpretowalność modelu, gdzie wkład każdego predyktora na wynik może być bezpośrednio pokazany. Autor osiągnął to poprzez modyfikację metody wzmocnienia gradientowego w wariancie składnikowym i dostosowanie jej do analizy przeżycia poprzez optymalizację funkcji straty błędu średniokwadratowego przeżycia o użycie zespołu drzew jako uczących bazowych i ograniczenie formy funkcjonalnej wyniku do sumy addytywnych nieliniowych funkcji pojedynczych predyktorów i ich interakcji parami. W pracy przeprowadzono szeroko zakrojone testy na rzeczywistych zbiorach danych, które potwierdziły wybory projektowe metody autora i wykazały jej wysoką wydajność na medycznych zbiorach danych.

Dokonany został przegląd właściwości interpretacyjnych metody, a następnie porównanie jej mocy predykcyjnej z innymi metodami uczenia maszynowego.

Obie tezy pomocnicze zostały potwierdzone w tradycyjnej analizie statystyki medycznej na zbiorze danych pacjentów po przeszczepie wątroby. Pokazano, że dodanie miar zmienności i trendu obserwacji biochemicznych prowadzi do modeli o lepszej mocy predykcyjnej, niż przy użyciu tylko miar statycznych, w tym uznanego modelu MELD.

Podkreślić należy także uniwersalność stosowanych metod pokazujących skuteczność proponowanych rozwiązań i ich przewagi wobec wcześniejszych metod znanych z literatury. Łatwość przejrzystej i prostej wizualizacji wpływu pojedynczych predyktorów i ich par interakcji na wynik jest kluczową właściwością proponowanej metody, która umożliwia intuicyjną interpretację wygenerowanych modeli. Z uznaniem należy odnotować korzystanie z międzynarodowych ogólnodostępnych baz danych. Proponowane rozwiązanie, przedstawione w tezie pracy, stanowi samodzielny i oryginalny dorobek Autora. Dzięki zastosowaniu nowoczesnych technik informatycznych, metod statystycznych, propozycji własnych algorytmów dla cząstkowych rozwiązań, praca reprezentuje wysoki poziom w skali międzynarodowej. Wnosi istotny wkład do literatury światowej dotyczącej jej tematyki — czyli metod analizy przeżycia może być pomocna dla **środowiska klinicznego**.

Struktura pracy oddaje wymagania wobec rozpraw doktorskich i właściwie grupuje w rozdziały i podrozdziały wykonane przez Doktoranta badania oraz osiągnięte wyniki. Treść podawana jest w zwięzły sposób, ale jest ona nadmiernie zwięzła, gdyż w wielu miejscach mogłaby być wzbogacona o autorskie podsumowanie, refleksje lub szczegóły stosowanych algorytmów. Brakuje wniosków oraz szczegółowych opisów proponowanych i zrealizowanych algorytmów.

Większość zagadnień jest prezentowana w poglądowy sposób i ilustrowana graficznie. Omawiane zagadnienia są przedstawione poprawnie pod względem merytorycznym.

Już na podstawie samej liczby pozycji literatury (142) można przypuszczać, że Autor dokonuje bardzo szczegółowego i obszernego przeglądu literatury adekwatnej do tematyki rozprawy. Przegląd literatury zawiera zasadnicze części: metody analizy danych, metody analizy statystycznej oraz przegląd metod ML. Niewątpliwie tak obszerny przegląd literatury światowej w zakresie stanu wiedzy o metodach analizy i ich praktycznych zastosowaniach świadczący o dużej wiedzy Autora. To pozwoliło mu zrealizować bardzo dojrzały proces projektowania i weryfikacji własnych rozwiązań mających na celu wykazanie zasadności zadanej tezy.

Tematyka rozprawy wpisuje się w aktualne światowe trendy badawcze dotyczące rozwoju metod i technik pozwalających na wykorzystanie gwałtownego rozwoju technologii, w tym ML.

Aktualność tematyki badawczej, szeroki zakres stosowanych metod i technik oraz weryfikacja na rzeczywistych danych w przyjętym scenariuszu zastosowania, pozwalają uznać, że jest to bardzo wartościowa praca dla rozwoju nauk technicznych i jej zastosowań.

Rozprawa stanowi propozycję innowacyjnego rozwiązania opracowanego dla rozwiązania trzech zasadniczych zagadnień występujących w specyficznym obszarze zastosowania. Praca w części eksperymentalnej - weryfikuje proponowane rozwiązania korzystając z dostępnych technologii pozyskiwania danych medycznych. Samo rozwiązanie polega na opracowaniu i zastosowaniu nowych algorytmów, zmodyfikowaniu istniejących bądź zastosowaniu uznanych technik ML. Odczuwam pewien niedosyt w opisie nowych algorytmów i ich weryfikacji.

Lektura rozprawy skłania do kilku uwag, tak ogólniejszej, jak i bardziej szczegółowej natury: Nie jest jasne dla jakich konfiguracji sprzętowych z konkretnymi pakietami/wersjami oprogramowania przygotowano i oceniano proponowaną metodę? Czy konsultowano wyniki z lekarzami specjalistami? Czy możliwe jest wykorzystanie opracowanego rozwiązania w rzeczywistych zastosowaniach poprzez zastosowanie opracowanej metody w systemie doradczym. Jak mógłby wyglądać proces wspierania decyzji w takim systemie?

Autor posiada wystarczający dorobek publikacyjny, obejmujący w 4 publikacje, w tym: 2 czasopismach z listy filadelfijskiej, 2 rozdziałach w książkach. Prace indeksowane są w bazach SCOPUS. Zebrały już 5 cytowań, a h-indeks Doktoranta wynosi 2.

Biorąc pod uwagę charakter pracy, nie dostrzegam żadnych zasadniczych wad rozprawy.

Uważam, że opiniowana rozprawa, pomimo zauważonych usterek stylistycznych, spełnia warunki stawiane przez obowiązującą ustawę o stopniach i tytule naukowym w określeniu do rozpraw doktorskich w dyscyplinie **informatyka techniczna i telekomunikacja** i wnioskuje jej dopuszczenie do publicznej obrony.