dr hab. Wojciech Rejchel, prof. UMK
Wydział Matematyki i Informatyki
Uniwersytet Mikołaja Kopernika

Toruń, 17.01.2025

Recenzja rozprawy doktorskiej mgr K. Kobylińskiej

„Metody wyjaśnialnego uczenia maszynowego dla danych tabelarycznych z przykładami zastosowań w medycynie"

Zwiększający się wpływ algorytmów uczenia maszynowego (z ang. *machine learning*, w skrócie ML) czy sztucznej inteligencji (z ang. *artificial intelligence*, AI) na wiele aspektów ludzkiego życia jest bezsporny. Medycyna jest jednym z głównych pól, na którym metody te mogą być bardzo pomocne, na przykład w poprawie diagnostyki bądź personalizacji leczenia. Naturalnym wymogiem jest, aby używane algorytmy były skuteczne i wydajne. W wielu zastosowaniach, zwłaszcza tych medycznych, równie ważną cechą jest *zrozumienie* procedur przez użytkownika (na przykład lekarza lub pacjenta). To ostatnie zagadnienie jest zasadniczym problemem w ocenianej rozprawie.

Dwa główne podejścia do problemu zrozumienia algorytmów to:

(a) konstrukcja algorytmów, które są z natury wytłumaczalne (z ang. *interpretable*), na przykład uogólnione modele liniowe lub ich wersje z ograniczeniami (chociażby opartymi na wiedzy eksperckiej),

(b) używanie złożonych (z natury niewytłumaczalnych) procedur, a następnie proponowanie metod, które potrafią wyjaśnić ich działanie. Podejście to zwykle określane jest skrótem *XAI* (z ang. *explainable* AI). Natomiast same złożone procedury często nazywane są *czarnymi skrzynkami* (z ang. *black boxes*, BB).

Autorka rozprawy skupiła się na drugim podejściu.

Omówienie rozprawy doktorskiej

Rozdział pierwszy jest wprowadzeniem do badanego problemu, a także zawarte w nim zostały cele pracy.

W rozdziale drugim omówiono główne metody i algorytmy ML, których użyto w rozprawie, na przykład są to drzewa losowe, lasy losowe czy procedury typu *boosting*. Następnie przestawiono podstawowe i dostępne w literaturze

metody XAI, jak wykresy częściowej zależności (z ang. *partial dependence plots*, PDP), permutacyjna ważność zmiennych, metoda wykorzystująca indeks Shapley'a czy profile *ceteris paribus*.

W rozdziale trzecim użyto powyższych metod do analizy trzech rzeczywistych zbiorów danych, które dotyczyły: przeżywalności pooperacyjnej osób chorujących na raka płuc, modeli przesiewowych raka płuc oraz przeżywalności pacjentów z sepsą przyjętych na oddziały intensywnej terapii.

W rozdziale czwartym wprowadzono dwa oryginalne narzędzia badawcze. Pierwszym jest miara PDI (z ang. *profile disparity index*), która porównuje profile procedur, bazując na ich monotoniczności. Drugim narzędziem jest algorytm *Rashomon_DETECT*, który służy do przeszukiwania, tak zwanego, zbioru Rashomon. Idea tego algorytmu polega na odejściu od standardowego poszukiwania (i analizy) jednego najlepszego algorytmu w kierunku badania kilku procedur, które są odpowiednio bliskie optymalnemu. Co więcej, w zbiorze tych kilku procedur poszukuje się algorytmów najbardziej zróżnicowanych. Moim zdaniem pomysł ten jest najciekawszą częścią rozprawy.

Ocena rozprawy i uwagi

Po pierwsze należy zauważyć oraz docenić owocną współpracę Autorki z wieloma ośrodkami naukowo-medycznymi w Polsce. Umiejętność nawiązania i utrzymania takich relacji (w szczególności przekonanie medyków do używania narzędzi ML w ich codziennej pracy) jest cechą bardzo cenną i dalece niepospolitą. Ponadto współpraca ta dotyczyła różnych chorób (na przykład białaczki, sepsy czy szpiczaka), co niewątpliwie wymagało przyswojenia sporej wiedzy medycznej.

Jak wspomniałem wcześniej, algorytm Rashomon_DETECT, a zwłaszcza idea za nim stojąca, jest, moim zdaniem, najważniejszym punktem rozprawy. Dlatego metoda ta powinna być głównym tematem pracy, a nie jej niewielką częścią. Ponadto w procedurze tej używa się uśredniania (po cechach oraz elementach zbioru Rashomon), co jest zrozumiałe. Zastanawiam się, czy użycie maksimum zamiast średniej (albo kombinacji maksimum i średniej) mogłoby prowadzić do poprawy rezultatów.

Następnie chciałbym omówić te fragmenty rozprawy, w których widzę najważniejsze niedociągnięcia bądź braki.

(1) Powyżej przytoczyłem dwa główne podejścia do problemu zrozumienia algorytmów. Rozprawa poświęcona jest drugiemu z nich, które krótko opisać można jako: *wyznacz BB*, a następnie *wyjaśnij BB*, czyli XAI. W pracy ewidentnie brakuje refleksji nad ryzykiem związanym z takim postępowaniem. Jednym z ważniejszych problemów jest interpretacja sytuacji, gdy wynik działania $BB + XAI$ jest niezgodny z wiedzą ekspercką. Przypadek ten może oznaczać, że BB działa *nieprawidłowo*. Jednakże może to również oznaczać, że BB działa prawidłowo, a XAI *myli się*. Odniosłem wrażenie, że w pracy ta druga możliwość jest pominięta, co jest mocno ryzykowne, wziąwszy pod uwagę oczywiste słabości metod XAI (jedną z nich omówię w punkcie drugim).

Powyższy problem nie jest jedynym zarzutem wobec XAI formułowanym w literaturze. Kolejne można znaleźć, na przykład, w pracy Cynthii Rudin pod wyrazistym tytułem *„Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead"*. C. Rudin jest niewątpliwie autorytetem w omawianej problematyce, co potwierdza fakt, że jej publikacje są aż dziesięciokrotnie cytowane w niniejszej rozprawie. Co prawda wspomniana powyżej praca nie jest cytowana, ale podobne stwierdzenia (wraz z argumentacją) można znaleźć w pracy oznaczonej numerem [99] w rozprawie, na przykład *„explainability and interpretability techniques are not alternative choices for many real problems, as the recent surveys often imply; one of them (XAI) can be dangerous for high-stakes decisions to a degree that the other is not"*. Przytoczyłem te stwierdzenia nie po to, aby negować przydatność metod XAI, ale chcąc zwrócić uwagę na ryzyko związane z ich używaniem. Refleksji tej nie znalazłem w rozprawie.

(2) Metody XAI mają dość liczne słabości, które nie są omówione w pracy. Dla przykładu przyjrzyjmy się PDP, czyli narzędziu kluczowemu, wielokrotnie używanemu w pracy. Cytując rozprawę, obiekt ten *„wyznacza średni profil odpowiedzi, czyli relację pomiędzy wartościami j-tej zmiennej a wartościami predykcji modelu"*. Zatem powinien być zdefiniowany jako *warunkowa* wartość oczekiwana

$$PD_f^j(z) = E\left[f(X_1, \ldots, X_{j-1}, z, X_{j+1}, \ldots, X_p | X_j = z\right].$$

Obiekt ten nie może być użyty w praktyce, gdyż wartość oczekiwana zależy od nieznanego rozkładu. Możnaby standardowo przybliżyć ją średnią, ale w tym przypadku jest to warunkowa wartość oczekiwana, więc należałoby użyć tylko danych z warstwy $\{X_j = z\}$. Naturalnie przybliżenie to byłoby niesatysfakcjonujące, gdyż danych z $\{X_j = z\}$ zwykle jest niewiele. Chcąc temu zaradzić, powszechnie używa się aproksymacji (2.10) z rozprawy, ktora bazuje na całym zbiorze danych. Jednakże jest to przybliżenie *bezwarunkowej* wartości oczekiwanej

$$E\left[f(X_1, \ldots, X_{j-1}, z, X_{j+1}, \ldots, X_p\right].$$

Dwa powyższe wzory są tożsame jedynie przy restrykcyjnym założeniu, że X_j jest niezależne od pozostałych cech, co trywializuje całe zagadnienie. Autorka nie komentuje tego faktu.

Powyższy problem pociąga kolejny, jakim jest użycie *nierealistycznych obserwacji* w aproksymacji (2.10), zwłaszcza gdy wartości z są relatywnie małe bądź duże. Dla przykładu przypuśćmy, że $j = 1$ oraz $X_1 - wzrost$, $X_2 - waga$, a także $z = 190$. Jeśli dane pochodzą ze standardowej populacji, to (2.10) oparte byłoby na, w większości nienaturalnych, obserwacjach postaci $(wzrost = 190, waga = prawdziwa\ waga\ osoby)$.

Podsumowując, uważam wyciąganie kategorycznych wniosków na podstawie PDP (zwłaszcza dla niestandardowych z) za postępowanie mocno ryzykowne. Mając to na uwadze, spójrzmy na tezę z podrozdziału 3.2 (strona 77) mówiącą, że wykresy PDP (Rysunek 3.8) oraz ich zestawienie z wiedzą medyczną (omówienie na stronach 70-71) ujawniły istotne błędy metod globalnych w procesie prog-

nozowania. Podchodząc ostrożnie do analizy PDP dla skrajnych wartości z, jedynie zachowanie zmiennej „qtyears" w modelu BACH uznaję za błędne.

(3) Jednym z osiągnięć rozdziału czwartego miało być wprowadzenie miary, która traktuje profile o podobnych kształtach jako bliskie sobie. Miara PDI nie w pełni realizuje ten cel, gdyż opiera się tylko na porównaniu znaków pochodnych dwóch profili, czyli ich przedziałów monotoniczności. Zatem miara PDI uzna profile za podobne, jeśli będą miały zbliżoną monotoniczność. Rozważmy dwa profile: jeden zadany funkcją wykładniczą, a drugi logarytmiczną. Kształty tych funkcji mocno różnią się, jednak ich miara PDI wynosi zero, gdyż obie są rosnące.

(4) Niektóre definicje podane są nieprecyzyjnie. Jest to zwłaszcza widoczne przy omawianiu metody SHAP (wzór (2.14)) czy modelu BACH (wzór (3.1) z opisem). Innym niedociągnięciem jest brak informacji, w jaki sposób wybrano parametry użytych procedur, na przykład liczba drzew czy ich głębokość w algorytmie lasów losowych w rodziale 3.1 albo ϵ w algorytmie Rashomon_DETECT w rozdziale 4.4.2.

Oczekuję, że Autorka dogłębnie odniesie się do powyższych uwag krytycznych podczas publicznej obrony rozprawy doktorskiej.

Konkluzja

Problem zrozumienia algorytmów jest zagadnieniem ważnym i intensywnie badanym. Uważam, że rozprawa zawiera oryginalne wyniki dotyczące metod XAI. Jednakże nie jest ona pozbawiona niedociągnięć i braków, co obniża moją ocenę.

Podsumowując, sądzę, że przedstawiona rozprawa doktorska spełnia ustawowe i zwyczajowe wymagania stawiane rozprawom doktorskim w dyscyplinie informatyka. Wnoszę o dopuszczenie mgr Katarzyny Kobylińskiej do dalszych etapów postępowania w sprawie nadania stopnia doktora.

Wojciech Rejchel