Streszczenie

Autor: **Bogdan Gulowaty**

Tytuł: Budowanie transparentnych modeli klasyfikacji

W miarę jak technologie Sztucznej Inteligencji (SI) i Uczenia Maszynowego rozwijają się i stają się coraz bardziej zintegrowane z codziennym życiem, potrzeba budowania przejrzystych i interpretowalnych modeli staje się coraz bardziej istotna. Systemy SI muszą być zrozumiałe, godne zaufania i etyczne, zwłaszcza te używane w kluczowych sektorach, takich jak opieka zdrowotna, finanse i systemy prawne. Rozwój Wyjaśnialnej Sztucznej Inteligencji (eXplainable Artificial Intelligence - XAI) ma na celu sprostanie tym wyzwaniom poprzez m.in. tworzenie bardziej przejrzystych i intepretowalnych modeli. Jednak większość prac dotyczących XAI koncentruje się na głębokich sieciach neuronowych, pozostawiając inne złożone modele, takie jak np. metody zespołowe bez rozwiązań w tym zakresie. Jednym z celów rozprawy jest wypełnienie tej luki poprzez opracowanie nowatorskich metod poprawy przejrzystości i interpretowalności klasyfikatorów zespołowych, przy jednoczesnym zachowaniu ich wysokiej jakości predykcji oraz budowę modeli klasyfikacji, które z góry zakładają swoją przejrzystosć. Główna hipoteza pracy stwierdza, że możliwe jest skonstruowanie takich przejrzystych lub wyjaśnialnych modeli, które będą działać równie dobrze jak modele typu "black box" w szerokim zakresie zadań predykcji danych. W pracy zaprezentowano trzy autorskie algorytmy wyjaśniania lub zastępowania złożonych modeli przez transparentne alternatywy: Non-overlapping Tree Ensemble (NOTE), Optimal Centroids oraz Quad Split. Przyjęte cele pracy obejmują:

- Opracowanie nowych algorytmów tworzących przejrzyste modele klasyfikacyjne.
- Ekstrakcję interpretowalnych modeli ze złożonych klasyfikatorów zespołowych, takich jak Random Forest (RF).
- Wykorzystanie metryk złożoności danych do oceny i poprawy wydajności przejrzystych modeli.
- Ocena skuteczności tych metod na zbiorach danych o różnej złożoności aby zapewnić, że modele są zarówno interpretowalne, jak i mają wysokie zdolności predykcyjne.

Głównym wkładem tej pracy jest zaproponowanie i ocena trzech nowatorskich algorytmów w dziedzinie XAI:

NOTE Metoda NOTE została zaprojektowana specjalnie do wyjaśniania zespołów modeli drzewiastych, takich jak RF. Tradycyjne modele zespołowe cechują się krokiem integracji predykcji z modeli będącymi członkami zespołu, co utrudnia interpretację zespołu. NOTE rozwiązuje ten problem, wybierając podzbiór drzew, które tworzą nienachodzące na siebie

obszary decyzyjne, upraszczając ogólny model. Proponowana metoda stosuje techniki analizy grafów w celu zidentyfikowania reguł pochodzących z drzew decyzyjnych w zespole, które można wykorzystać do zdefiniowania nienachodzących na siebie obszarów, zapewniając przy tym jak najlepsze możliwości predykcyjne modelu. Obszary pokryte przez wybrane reguły są przypisywane do drzew decyzyjnych, które są trenowane i łączone w komitet, tworząc przejrzystą wersję oryginalnego modelu, umożliwiając użytkownikom łatwiejsze zrozumienie indywidualnych predykcji. Wyniki eksperymentów wykazały, że metoda NOTE poprawia interpretowalność, nie tracąc przy tym znacząco na wydajności predykcyjnej. Na zbiorach danych o umiarkowanej złożoności metoda NOTE osiągnęła wyniki porównywalne z oryginalnym Random Forest, jednocześnie dostarczając dużo precyzyjniejszych informacji o procesie podejmowania decyzji.

Optimal Centroids Algorytm Optimal Centroids koncentruje się na tworzeniu interpretowalnych modeli klasyfikacyjnych poprzez podział przestrzeni cech na regiony w oparciu o centroidy i klasyfikator 1-NN. W tej metodzie przestrzeń cech jest podzielona zgodnie z centroidami, które są optymalizowane za pomocą algorytmów ewolucyjnych, takich jak algorytm genetyczny. Pozycje centroidów są oceniane w celu znalezienia ich najlepszej dystrybucji, która maksymalizyje dokładność modelu. W ramach kroku ewaluacji trenowane są przejrzyste modele dla przestrzeni decyzyjnych wyznaczonych przez centroidy. Tak wytrenowany model cechuje się wysoką interpretowalnością i dobrymi zdolnościami predykcyjnymi.

Quad Split Algorytm Quad Split to nowatorska metoda tworzenia przejrzystych modeli klasyfikacyjnych, zaprojektowana z myślą o konkurowaniu z bardziej złożonymi modelami poprzez tworzenie interpretowalnych granic decyzyjnych. Algorytm ten zakłada, że zastosowanie wielu wyspecjalizowanych modeli do części zbioru danych przyniesie lepsze wyniki niż jeden model globalny. Aby to osiągnąć, Quad Split rekursywnie dzieli przestrzeń decyzyjną w oparciu o wartości cech wyodrębnionych ze zbioru treningowego. Wartości te służą jako punkty podziału. Algorytm ocenia te punkty, aby znaleźć taki, który minimalizuje złożoność danych po obu stronach podziału. Po osiągnięciu kryteriów zatrzymania w każdym podziale tworzony jest interpretowalny model, co ostatecznie prowadzi do utworzenia zespołu przejrzystych modeli, który można sprowadzić do listy prostych reguł.

Ocena eksperymentalna W ramach pracy przeprowadzono szeroki zakres badań eksperymentalnych trzech proponowanych metod, używając 16 zbiorów danych do zadania binarnej klasyfikacji o zróżnicowanej złożoności. Zbiory te obejmowały zarówno przykłady rzeczywiste, jak i syntetyczne, dotyczyczące różnych dziedzin i wyzwań, takich jak niezbalansowane rozkłady klas, przestrzenie o wysokiej wymiarowości oraz szum danych. Eksperymenty brały pod uwagę ocenę złożoności zbiorów danych w ramach różnych kategorii, takich jak cechy, wymiarowość i liniowość. Eksperymenty koncentrowały się na kilku kluczowych wskaźnikach oceny:

- Zdolności predykcyjne: Wyniki każdego modelu, mierzone przez zbalansowaną dokładność, F1-score i średnią geometryczną, zostały porównane ze standardowymi klasyfikatorami, takimi jak Random Forest, drzewa decyzyjne oraz modele regułowe, z naciskiem na ogólną dokładność predykcji.
- Złożoność modelu: Złożoność podstawowych i proponowanych modeli została skwantyfikowana i oceniona w stosunku do siebie.

- Skuteczność jako metody wyjaśniania modeli: Oceniono zdolność proponowanych algorytmów do wyjaśniania modeli typu "czarna skrzynka", takich jak Random Forest. Eksperymenty oceniły, jak dobrze proponowane metody mogą pełnić rolę wyjaśniających modeli zastępczych dla tych bardziej złożonych.

Wszystkie powyższe cechy zostały ocenione w kontekście zmieniających się parametrów wewnętrznych algorytmów oraz różnych złożoności zbiorów danych, zdefiniowanych przez wspomniane metryki złożoności.

Wnioski i przyszłe kierunki badań Praca wnosi istotny wkład do rozwoju AI, pokazując, że można budować przejrzyste i wyjaśnialne modele bez poświęcania ich zdolności predykcyjnych. Proponowane metody — NOTE, Optimal Centroids i Quad Split — oferują elastyczne i interpretowalne alternatywy dla tradycyjnych modeli zespołowych typu "czarna skrzynka". Główne osiągnięcia pracy obejmują:

- Opracowanie dwóch nowatorskich algorytmów, które generują przejrzyste modele przy zachowaniu konkurencyjnych zdolności predykcji i generalizacji.
- Opracowanie algorytmu wyjaśniającego dla komitetów drzew decyzyjnych, który jest konkurencyjny wobec algorytmu RF.
- Szeroką ocenę eksperymentalną, pokazującą, że proponowane metody mogą wyjaśniać, a w niektórych przypadkach zastępować modele typu "czarna skrzynka" przejrzystymi alternatywami.
- Wykazanie, że metryki złożoności danych odgrywają ważną rolę w określaniu, kiedy przejrzyste modele są najbardziej efektywne.

Trzy zaproponowane metody wnoszą znaczące ulepszenia w zakresie interpretowalności, co czyni je odpowiednimi do zastosowań, w których zaufanie i przejrzystość są kluczowe, takich jak opieka zdrowotna, finanse czy zastosowanie w problemach prawnych. Oprócz powyższego, poczyniono kilka obserwacji, takich jak:

- Oceniane modele klasyfikacyjne zachowują się znacząco inaczej, gdy są stosowane do zbiorów danych o różnych właściwościach złożoności.
- Wewnętrzna złożoność modeli nie zawsze koreluje ze złożonością zbiorów danych.
- Oceniane metody wyjaśniające, które wykorzystywały wiedzę wyodrębnioną z złożonych modeli, działały lepiej niż z natury przejrzyste modele, gdy były stosowane jako modele wyjaśniające.

Przyszłe badania mogłyby skupić się na dalszym udoskonalaniu proponowanych metod, na przykład poprzez dostrajanie parametrów algorytmu genetycznego w metodzie Optimal Centroids lub poszukiwanie lepszych metryk oceny dla NOTE. Dodatkowo uzasadnione jest sprawdzenie zastosowanie algorytmów w szerszym zakresie złożonych metod typu "czarna skrzynka", takich jak sieci neuronowe. Obserwacje poczynione w pracy mogą również pomóc w opracowaniu meta-algorytmów opartych na metrykach złożoności zbiorów danych. Praca wnosi znaczący wkład do dziedziny XAI, wykazując, że można budować przejrzyste modele bez poświęcania ich zdolności predykcyjnych.