DOI: 10.24425/academiaPAN.2023.147438

ALGORYTMY STOSOWANE W MODELOWANIU LASÓW

Uczenie maszynowe zmieniło podejście do analizy danych i jest odpowiedzią na postępujący przyrost objętości i złożoności danych podlegających analizie.

Łukasz Pawlik

Wydział Nauk Przyrodniczych
Uniwersytet Śląski w Katowicach

Marcin K. Dyderski

Instytut Dendrologii Polskiej Akademii Nauk
w Kórniku

dr hab. Łukasz Pawlik, prof. UŚ

Jest geografem i geomorfologiem, profesorem Uniwersytetu Śląskiego pracującym w Instytucie Nauk o Ziemi UŚ. Prowadzi badania nad wpływem czynników biotycznych i zaburzeń naturalnych na procesy geomorfologiczne, formy rzeźby i ewolucję gleb górskich.

lukasz.pawlik@us.edu.pl

dr hab. Marcin K. Dyderski, prof. PAN

Jest profesorem Instytutu Dendrologii PAN. Prowadzi badania dotyczące reakcji roślin na działalność człowieka, obejmujące wpływ górnictwa, gospodarki leśnej, zmian klimatycznych i inwazji biologicznych, w szczególności inwazyjnych gatunków drzew.

mdyderski@man.poznan.pl

Uczenie maszynowe jest częścią bardzo ambitnej i różnie ocenianej koncepcji zorientowanej na rozwój sztucznej inteligencji, która jest wykorzystywana w wielu dziedzinach nauki, przemyśle i usługach. Jest to rezultat wielkiego marzenia człowieka o stworzeniu maszyny-systemu zdolnego do uczenia się i rozpoznawania wzorców, tak jak robi to najdoskonalsza „maszyna" będąca efektem ewolucji, czyli ludzki mózg. Odkrywanie wiedzy z danych przyświecało pierwszym systematycznym obserwacjom zjawisk przyrodniczych. W ten sposób mieszkańcy starożytnego Egiptu około 3,5 tys. lat p.n.e. zdefiniowali rytm wezbrań Nilu, by zaplanować okresy zasiewów i zbierania plonów. Wykorzystali do tego uproszczony schemat matematyczny danego zjawiska, czyli model. Z czasem ilość danych zaczęła gwałtownie rosnąć, a dostępne metody statystyczne, często o rygorystycznych założeniach, nie oferowały oczekiwanych rozwiązań. Obecnie, w erze wielowymiarowego „potoku danych" i Big Data, ich wyjaśnianie (*data mining*) jest wspomagane przez zaawansowane algorytmy i narzędzia, tj. centra superkomputerowe oraz klastry obliczeniowe. Ta specyficzna walka z danymi, jak się wydaje, jest dopiero na początkowym etapie, a napływ nowych informacji oraz oczekiwania odbiorców mogą być kolejnymi punktami zapalnymi rozwoju uczenia maszynowego. Jednocześnie pojawiają się krytyczne opinie, które są mniej lub bardziej medialne. Jednym z poważniejszych głosów w toczącej się dyskusji nad zagrożeniami ze strony sztucznej inteligencji jest książka Cathy O'Neil *Broń matematycznej zagłady*, w której autorka stwierdza m.in.: „Tworząc model, dokonujemy wyboru tego, co wydaje się nam wystarczająco ważne, by to uwzględnić. Upraszczamy świat". Jak się okazuje, takie uproszczenia mogą mieć w pewnych przypadkach negatywne konsekwencje, np. w postaci dyskryminacji Afroamerykanów w procedurze aplikowania o kredyt hipoteczny czy mieszkańców Podhala przy staraniu się o wizy do USA.

Idea stojąca za nauczeniem pewnego systemu rozpoznawania schematów pojawiła się już bardzo wcze-

śnie i wiązała się z obserwacjami astronomicznymi. System uczący się został zdefiniowany w 1997 roku przez Toma Mitchella w książce *Machine Learning* jako ten, który udoskonala się wraz z doświadczeniem. Z kolei Marcin Szeliga w *Praktycznym uczeniu maszynowym* dodaje, że uczenie maszynowe polega na uczeniu maszyn na przykładach.

W biologii w 1958 roku pojawiło się pojęcie perceptronu naśladującego działanie neuronów w mózgu, a w 2001 roku pojawiła się klasyczna w swojej wymowie publikacja Leo Breimana *Random Forests* prezentująca, jak za pomocą drzew decyzyjnych wyizolować model klasyfikacyjny. Ponieważ w przyrodzie większość zjawisk funkcjonuje na zasadzie naczyń połączonych, zakłada się, że pewne zjawisko lub właściwość środowiska przyrodniczego (Y) można wyjaśnić za pomocą zbioru predyktorów (x_1, …., x_n), które pełnią funkcję zmiennych niezależnych. Nie wszystkie predyktory wyjaśniają badane zjawisko lub cechę w tym samym zakresie, a ich siła (wpływ) jest analizowana w trakcie pierwszego etapu modelowania. Kluczowym elementem całej układanki są dane historyczne, na których bazie system uczący się zdobywa wiedzę o schematach. Następnie ją aplikuje w postaci modelu do przestrzeni niepróbkowanej (np. przestrzeni geograficznej) części populacji czy zjawisk dynamicznych, prognozując ich przyszłe stany.

Wielowymiarowe zbiory danych zawierające zmienną zależną i dużą liczbę predyktorów o często niejednolitym formacie (dane ciągłe lub kategoryczne) i odmiennych rozkładach prawdopodobieństwa wymagają szczególnego podejścia. Zadaniem algorytmu uczenia maszynowego jest stworzenie modelu, który pozwala na predykcję. Max Kuhn i Kjell Johnson w swojej książce *Applied Predictive Modeling* (Zastosowanie modeli predykcyjnych) zawierają sens tego terminu w następujący sposób: „Model lub modelowanie (...) jest to proces tworzenia narzędzia matematycznego (...), które pozwala na precyzyjną predykcję". Z kolei Brad Boehmke i Brandon Greenwell w *Hands-On Machine Learning with R* (Praktyczne uczenie maszynowe z R) zwracają uwagę, że ważną cechą procesu uczenia maszynowego jest proces iteracyjny, bazujący na podejściu heurystycznym. Może się zdarzyć, że niewiele wiemy na temat analizowanego zjawiska i zazwyczaj opieramy swoje działania na niepełnych danych. Nie wiemy, która metoda uczenia maszynowego najlepiej odda rzeczywisty schemat badanego zjawiska lub stan środowiska naturalnego. Stąd potrzeba zastosowania, ewaluacji, zmodyfikowania metody lub danych i ponownego tworzenia modelu (treningu) na tym samym zbiorze danych. W wielu przypadkach takie podejście daje najlepszy pożądany efekt

– model o pewnym stopniu uogólnienia (nieprzetrenowany), możliwy do zastosowania dla wielu różnych zbiorów danych testowych.

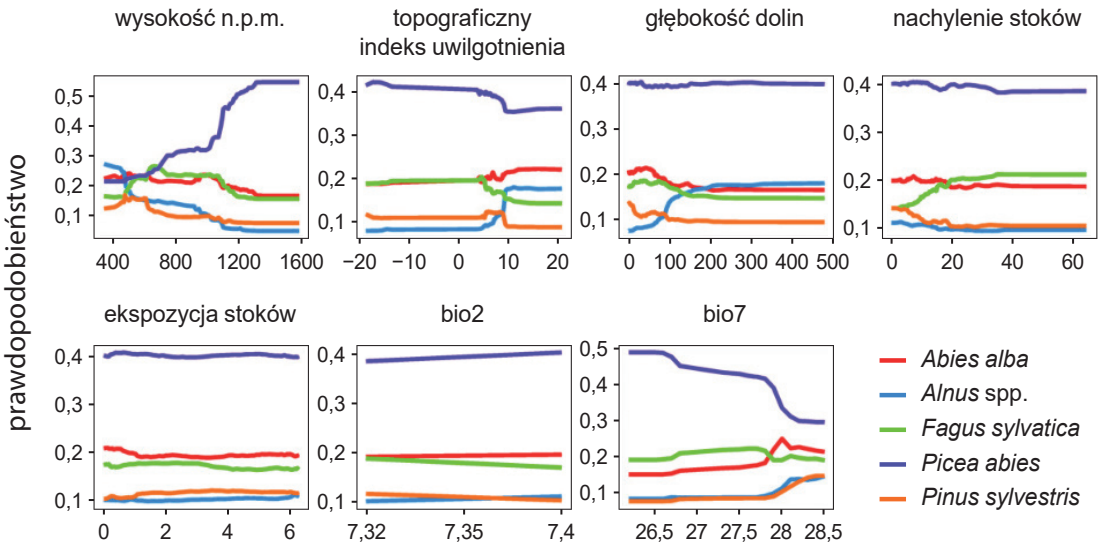
Dane

Bogactwo różnego rodzaju danych sprawia, że już ich wstępna ocena i przygotowanie stanowi ważny etap modelowania. Ocenia się, że około 80 proc. czasu poświęconego na analizę zajmuje samo przygotowanie danych. Na tym etapie należy wziąć pod uwagę zasadę GIGO, czyli „garbage in, garbage out" (śmieci na wejściu, śmieci na wyjściu), podkreślającą, że błędne, niekompletne i bardzo małe zbiory danych mogą prowadzić do błędnego wnioskowania i konkluzji. Dane obserwacyjne zbierane za pomocą metod konwencjonalnych – instrumentów pomiarowych – często stanowiły zbiory o małych woluminach i były obarczone błędami wynikającymi z niedokładności metody pomiarowej (rozdzielczości przestrzennej i czasowej), instrumentu pomiarowego lub jego awarii. Ponieważ nie da się pomierzyć wszystkiego, wszelkie analizy prowadzą do pewnego uproszczenia, którego finalną postacią jest model. Przez analogię można w tym miejscu przywołać najprostszą kartograficzną definicję mapy, mówiącą, że mapa jest modelem rzeczywistości.

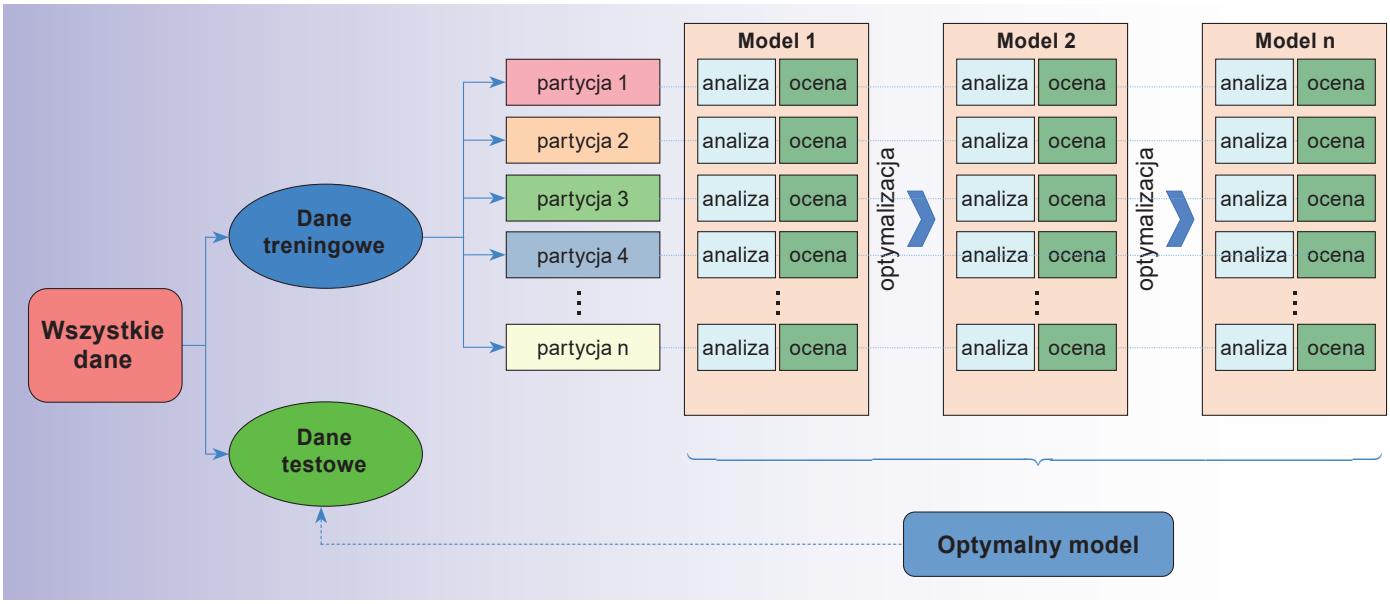
Wraz z pojawieniem się systematycznych pomiarów laboratoryjnych, zobrazowań satelitarnych i np. automatycznych pomiarów meteorologicznych, wielkie woluminy danych okazały się szansą zrobienia kroku milowego w lepszym poznaniu złożoności świata fizycznego i biologicznego. Jednak odczytanie informacji w gąszczu danych wymagało nowych technik obliczeniowych, większej objętości baz danych i szybszych procesorów. Obecnie użytkownicy internetu sami produkują dane (zostawiają cyfrowy ślad), które następnie wchodzą do modelu decydującego o tym, jakie np. reklamy są wyświetlane na ekranie ich komputerów. Systemy odpowiedzialne za skrzynki e-mailowe są uczone, jak rozpoznawać wiadomość typu spam, a na bramkach autostradowych są rozpoznawane tablice rejestracyjne samochodu, co pozwala – lub nie – na dalszy przejazd.

Uczenie nadzorowane

Jedną z ważniejszych cech zbioru danych, która warunkuje możliwość zastosowania pewnej części metod uczenia maszynowego, jest to, czy obserwacje zawierają informacje o zmiennej zależnej, czyli odpowiedź modelu (*target variable,* czyli zmienna docelowa) przy określonym zestawie predyktorów. Jeżeli zbiór danych zawiera etykiety (w modelu klasyfikacyjnym, w których zmienna zależna może mieć charakter binarny, np. tak – nie, 0 – 1, zniszczenie – brak zniszczeń) lub konkretne wartości liczbowe (w modelu regresyjnym, np. szacującym biomasę drzewostanów), można zastosować uczenie nadzorowane, gdzie na wejściu przekazujemy informację o zmiennej Y. W ten sposób „nadzorujemy" pewien oczekiwany wynik modelu w efekcie zastosowania danego algorytmu, np. lasów losowych (*random forests*). W tym przypadku warto wspomnieć o pewnym ograniczeniu metody, która może wynikać z samej natury zjawiska. Jeżeli jest analizowany wzrost drzew zależny od temperatury i opadów, to opierając się na danych tylko ze strefy klimatu umiarkowanego, model nie będzie mógł być zastosowany do analizy lasów w innej strefie klimatycznej, ponieważ został wytrenowany na podstawie obserwacji w pewnym zakresie wartości oraz dla pewnych kategorii drzew, niewystępujących w strefie, dla której użytkownik planuje zastosować model. Wskazujemy tu na ważną cechę modeli regionalnych. Przykładowo w analizie przestrzennej na problem ten wskazali Meyer i Pebesma, którzy proponują w artykule z 2021 roku



Wykresy cząstkowe pokazujące uśrednione prawdopodobieństwo wystąpienia analizowanego gatunku drzewa. Bio2 – dobowy zakres temperatury uśredniony dla roku, bio7 – różnica między maksymalną temperaturą najcieplejszego miesiąca a minimalną temperaturą najchłodniejszego miesiąca w roku



Schemat uczenia na danych treningowych w iteracyjnym procesie modelowania i oceny aż do wypracowania optymalnego modelu (na podstawie Boehmke i Greenwell, 2020)

Predicting into unknown space? Estimating the area of applicability of spatial prediction models walidację modelu za pomocą metody *area of applicability* (AOA – obszar zastosowania). W wielkim skrócie jest to obszar, w którym model może nauczyć się relacji między zmiennymi w oparciu o dane treningowe i oszacowana w trakcie walidacji krzyżowej jakość modelu jest utrzymana na pewnym akceptowalnym poziomie.

Ważną cechą walidacji modelu nadzorowanego jest wydzielenie pewnej części danych ze zbioru głównego, np. 25 proc. obserwacji, i pozostawienia ich jako zbioru testowego, który nie jest używany do modelowania (wytrenowania). W ten sposób utrzymujemy kontrolę nad jakością modelu. Warto pamiętać, że model może ulec przetrenowaniu, tzn. posiada dużą moc predykcyjną dla zbioru o podobnych charakterystykach (podzbioru danych), ale staje się bezużyteczny przy zastosowaniu dla zupełnie nowych danych.

Zastosowanie algorytmów

Algorytm lasów losowych to jedna z najpopularniejszych metod uczenia maszynowego opierająca się na wielu drzewach decyzyjnych lub klasyfikacyjnych. W trakcie budowania drzewa jest wybierana losowa próbka *m* predyktorów (zmiennych niezależnych). Gdy jest poszukiwany model klasyfikacyjny, wartość ta równa się pierwiastkowi kwadratowemu z liczby predyktorów. Użycie niewielu predyktorów pomaga w obejściu problemu ich współliniowości (korelacji), ponieważ są one wybierane losowo w trakcie budowania np. 1000 drzew klasyfikacyjnych. Dzięki takiej procedurze algorytm lasów losowych dobrze radzi sobie nawet z danymi silnie skorelowanymi, ponieważ na końcu wynik jest uśredniany dla wielu drzew.

Metoda ta została wykorzystana m.in. do modelowania przestrzennego rozkładu głównych gatunków drzew i ich biomasy w parkach narodowych południowej Polski oraz modelowania zniszczeń spowodowanych przez orkan Klaus w 2009 roku w lasach południowo-zachodniej Francji.

Do wyznaczenia wzorców występowania poszczególnych gatunków drzew w lasach górskich wykorzystano dane z opisów drzewostanów z pięciu parków narodowych oraz mapy przedstawiające klimat i cechy geomorfometryczne tych obszarów. Z kolei żeby określić dominujący gatunek drzewostanu, zastosowano model klasyfikacyjny, przypisujący każdej obserwacji (drzewostanowi) jeden z pięciu gatunków. W tym modelu drzewa klasyfikacyjne decydowały w oparciu o zestaw danych, wskazując prawdopodobieństwo, że w danym miejscu będzie dominował konkretny gatunek drzewa. Wykazaliśmy, że dla poszczególnych gatunków różne czynniki determinowały ich występowanie. Na przykład prawdopodobieństwo występowania świerka było związane głównie z wysokością nad poziomem morza, a sosny – z wystawą i nachyleniem. Zastosowanie tego modelu pozwoliło więc wyjaśnić, które czynniki są najważniejsze dla występowania gatunku. Dodatkowo wizualizacja odpowiedzi modelu zakładająca stały poziom wszystkich zmiennych z wyjątkiem jednej pozwoliła na symulację zmian warunków środowiskowych, wskazując, jak modyfikacja poszczególnych czynników wpływa na badane gatunki. Z kolei zastosowanie modelu regresyjnego pozwoliło wnioskować, jak jednostkowa zmiana danego predyktora (np. wystawy) wpływa na biomasę drzewostanu. W ten sposób nasze modele pozwoliły przewidzieć, jak zmiany klimatu czy topografii mogą wpływać na zdolność drzew do akumulacji węgla i przeciwdziałania zmianom klimatycznym. ■

Chcesz wiedzieć więcej?

Dyderski M.K., Pawlik Ł., *Spatial distribution of tree species in mountain national parks depends on geomorphology and climate*, „Forest Ecology and Management" 2020, doi.org/10.1016/j.foreco.2020.118366

Pawlik Ł., Godziek J., Zawolik Ł., *Forest damage by extra-tropical cyclone Klaus – modelling and prediction*, „Forests" 2022, doi: 10.3390/f13121991

Pawlik Ł., Harrison S.P., *Modelling and prediction of wind damage in forest ecosystems of the Sudety Mountains*, „Science of the Total Environment" 2022, doi.org/10.1016/j.scitotenv.2021.151972