

## STRESZCZENIE ROZPRAWY DOKTORSKIEJ

### PREDYKCJA CHARAKTERYSTYK MAŁYCH OBSZARÓW WSPOMAGANA METODAMI UCZENIA MASZYNOWEGO

Autor: mgr Adam Chwila

Promotor: dr hab. Tomasz Żądło, prof. UE

Praca wpisuje się w podejmowane w literaturze w ostatnim czasie próby integracji podejścia statystyki małych obszarów oraz podejścia uczenia maszynowego do problemu predykcji opartej na technikach regresyjnych. Statystyka małych obszarów bazuje na technikach wykorzystywanych w ramach metody reprezentacyjnej, która jest działem statystyki matematycznej. Techniki uczenia maszynowego nie są jednoznacznie definiowane w literaturze, natomiast w swojej istocie skupiają się na algorytmizacji kolejnych etapów rozpatrywanego problemu wnioskowania na podstawie danych. Obecnie w ramach statystyki publicznej oraz metody reprezentacyjnej można zaobserwować stopniową transformację obowiązujących paradygmatów. Pierwsza transformacja dotyczy przejścia od paradygmatu badań losowych do paradygmatu stosowania mieszanych źródeł danych, w tym dużych, rozproszonych zbiorów (ang. *big data*). Druga transformacja dotyczy zmiany paradygmatu statystyki publicznej z kultury modelowania danych na kulturę modelowania algorytmicznego. W opinii autora kierunkiem pozwalającym na podjęcie obecnych wyzwań dotyczących statystyki małych obszarów jest m.in. integracja ram statystyki klasycznej oraz uczenia maszynowego.

Integracja obu podejść wymaga m.in. dostosowania predyktorów i estymatorów wykorzystywanych obecnie w statystyce małych obszarów do modeli uczenia maszynowego oraz zaproponowanie metod umożliwiających ocenę dokładności predykcji. W przypadku oceny dokładności predykcji ważnym problemem jest możliwość porównania dokładności predyktorów wykorzystujących różne podejścia w praktyce tj. na podstawie danej próby (bez konieczności prowadzenia badań symulacyjnych). W szczególności w tym zakresie można rozważać ocenę dokładności predykcji różnych predyktorów, ale przy założeniu klasycznych modeli stosowanych obecnie w statystyce małych obszarów (np. liniowych modeli mieszanych), co pozwala na ocenę i porównania dokładności predykcji zgodnie z podejściem obecnie stosowanym w statystyce małych obszarów. Ponadto ocena i porównania dokładności predykcji predyktorów wykorzystujących różne podejścia mogą być prowadzone przy

założeniu dowolnego modelu uczenia maszynowego, co zaproponowano w pracy. Problem integracji obejmuje również omówienie roli, jaką mogą pełnić hiperparametry metod uczenia maszynowego w procesie szacowania charakterystyk podpopulacji. Hiperparametry modeli uczenia maszynowego mają znaczny wpływ zarówno na końcową postać architektury modelu (np. liczba drzew lasu losowego, liczba neuronów w sieci neuronowej), jak i na proces wyznaczania wartości parametrów, w tym na postać optymalizowanej funkcji celu (np. poprzez stosowanie regularyzacji).

Głównym celem pracy jest porównanie istniejących parametrycznych i nieparametrycznych metod szacowania charakterystyk małych obszarów oraz zaproponowanie modyfikacji, które pozwolą na zwiększenie dokładności oszacowań. W szczególności skupiono się na roli, jaką mogą pełnić hiperparametry metod uczenia maszynowego w procesie szacowania charakterystyk podpopulacji i na ocenie dokładności predykcji *ex ante* niezależnie od klasy, do której należą rozpatrywane predyktory. W pracy zastosowano następujące metody badawcze: krytyczna analiza literatury przedmiotu, metody statystyczno-ekonometryczne oraz metody symulacji komputerowych. Badania empiryczne i symulacyjne zostały przeprowadzone przez autora pracy w języku R.

Struktura pracy została podzielona na cztery rozdziały. W rozdziale pierwszym zostały zaprezentowane definicje małego obszaru i uczenia maszynowego oraz podstawowe pojęcia wykorzystywane w ich ramach. Został omówiony problem wykorzystania metod uczenia maszynowego w ramach statystyki małych obszarów oraz wyzwania, jakie stoją przed współczesnymi badaczami podejmującymi tę tematykę. Zostały przedstawione argumenty przemawiające za wykorzystaniem metod uczenia maszynowego w statystyce małych obszarów z uwzględnieniem typowych problemów praktycznych, z którymi mają do czynienia badacze np. problemu relatywnie niewielkiej liczby danych. W ramach rozdziału pierwszego została zaproponowana nowa autorska procedura doboru hiperparametrów modeli uczenia maszynowego, która uwzględnia specyfikę szacowanych charakterystyk małych obszarów. Dzięki zaproponowanemu rozwiązaniu możliwe jest osiągnięcie dokładniejszych szacunków charakterystyk w domenach niż w przypadku standardowych metod doboru hiperparametrów, co zostało przedstawione w rozdziale poświęconym badaniom empirycznym. Rozdział pierwszy zamyka przegląd prób wykorzystywania metod uczenia maszynowego w statystyce małych obszarów, ze szczególnym uwzględnieniem urzędów statystycznych i innych instytucji zajmujących się statystyką publiczną.

W rozdziale drugim zostały przedstawione modele statystyczne powszechnie wykorzystywane w ujęciu klasycznym statystyki małych obszarów oraz wybrane modele

nieparametryczne, wykorzystywane w ramach uczenia maszynowego. W szczególności zostały opisane modele liniowe oraz liniowe modele mieszane powszechnie wykorzystywane w statystyce małych obszarów. Następnie przedstawiono wybrane modele uczenia maszynowego takie jak modele liniowe z regularyzacją, wieloraka regresja adaptacyjna (MARS), regresja z wykorzystaniem wektorów nośnych (SVR), drzewa decyzyjne, lasy losowe, modele wzmacniane gradientowo, czy sieci neuronowe. Została przedstawiona charakterystyka omawianych metod, wykorzystywane hiperparametry oraz parametry, implementacje komputerowe ułatwiające wykorzystanie praktyczne omawianych metod, a także zalety i wady omawianych modeli.

W rozdziale trzecim zostały przedstawione predyktory wykorzystywane w ramach statystyki małych obszarów. Został również rozwinięty problem wykorzystania wybranych predyktorów dla modeli uczenia maszynowego. Następnie omówiono techniki łączenia wielu predyktorów, co jest często rozważanym podejściem w ramach uczenia maszynowego. W ramach rozdziału trzeciego przedstawiono kilka nowych propozycji. Została rozwinięta autorska propozycja estymatorów wspomaganych metodami uczenia maszynowego oraz wykorzystanie predyktorów plug-in dla modeli uczenia maszynowego. Została również zaproponowana autorska metoda łączenia wielu predyktorów opartych o algorytmy stochastyczne, skonstruowana z uwzględnieniem problemów występujących w ramach predykcji charakterystyk podpopulacji. Następnie została zaproponowana autorska procedura bootstrap, pozwalająca na ocenę dokładności *ex ante* predykcji w przypadku założenia modelu uczenia maszynowego jako modelu nadpopulacji. Procedura ta wykorzystuje stosowanie *k*-krotnej walidacji krzyżowej oraz uwzględnia wpływ procesu wyboru hiperparametrów na dokładność prognoz. Procedura bootstrap została zaproponowana w dwóch wariantach różniących się czasem obliczeń komputerowych. Rozdział trzeci zamyka omówienie problemu interpretowalności wyników uzyskiwanych z wykorzystaniem modeli uczenia maszynowego.

Rozdział czwarty został poświęcony badaniom empirycznym przeprowadzonym dla problemu szacowania charakterystyk małych obszarów w podpopulacjach, gdzie zmienną badaną jest cena domów w USA. Dane, które są wykorzystywane w badaniach empirycznych, są przykładem typowych danych, z których korzystają praktycy, zajmujący się problemem predykcji tzn. dostępny zbiór danych populacyjnych cechuje się ograniczeniami, jakimi są przykładowo brak istotnych zmiennych, które pozwoliłyby na precyzyjne modelowanie charakterystyk podpopulacji. Zostały zaprezentowane wyniki analizy dokładności *ex post* oszacowań uzyskane z wykorzystaniem technik statystyki małych obszarów oraz uczenia maszynowego. Wyniki analiz *ex post* dokładności prognoz wskazują, że pomimo użycia

zmiennych objaśniających charakteryzujących się liniową zależnością ze zmienną badaną dla 69% szacowanych charakterystyk w poszczególnych domenach (25 na 36) niższymi błędami charakteryzują się oceny otrzymane za pomocą modeli uczenia maszynowego. Ponadto przeprowadzone badania na danych rzeczywistych wykazały, że zaproponowane dostosowanie sposobu doboru hiperparametrów modelu uczenia maszynowego do rozważanych charakterystyk małych obszarów może mieć w określonych wypadkach znacząco większy wpływ na poprawę dokładności prognoz, niż zysk, który można osiągnąć, rozważając metody uczenia maszynowego, jednak nieuwzględniające specyfiki problemów małych obszarów. Zostały również oszacowane wartości miernika dokładności *ex ante* predykcji dla przypadku założenia jako modelu nadpopulacji modelu regresji z wykorzystaniem wektorów nośnych, co pozwala na praktyczne wykorzystanie autorskich technik oceny dokładności predykcji z wykorzystaniem metody bootstrap zaproponowanych w rozdziale trzecim. Zarówno w przypadku analiz *ex post* jak i *ex ante* została wykorzystana autorska metoda doboru hiperparametrów modeli uczenia maszynowego, przedstawiona w rozdziale pierwszym, która w wielu przypadkach pozwoliła na znaczną poprawę dokładności predykcji charakterystyk małych obszarów. Zostało również przeprowadzone badanie symulacyjne w ramach podejścia modelowego statystyki małych obszarów. Badania symulacyjne potwierdziły dobre własności zaproponowanej w podrozdziale 1.4 metody doboru hiperparametrów oraz potwierdziły, że stosowanie technik uczenia maszynowego może poprawić dokładność prognoz nawet w przypadku korzystnych warunków dla stosowania modeli liniowych. Uzyskane symulacyjnie wartości miernika RMSE dokładności *ex ante* predykcji dla poszczególnych charakterystyk w domenach wykazały, że na 22 z 36 przypadków najniższe wartości wystąpiły dla predyktorów korzystających z rozwiązań uczenia maszynowego, w tym aż dla 13 przypadków najlepszy wynik został osiągnięty poprzez zastosowanie autorskiej metody doboru hiperparametrów.

Całość pracy wieńczy zakończenie, w którym znajduje się synteza wniosków z przeprowadzonych rozważań i badań.

Autor podkreśla, że na uzyskane wyniki ma wpływ specyfika prowadzonych badań. Niemniej jednak wykorzystanie metod uczenia maszynowego, w tym autorskich propozycji zaprezentowanych w wymiarze praktycznym w rozdziale czwartym, może pozwolić na znaczącą poprawę dokładności oszacowań. Jednocześnie warto podkreślić, że w erze rozważania dużych, rozproszonych zbiorów danych korzystanie z technik nieparametrycznych, niewymagających spełnienia dodatkowych założeń i odpowiedniego przygotowania danych

wejściowych może znacząco wpłynąć na efektywność prowadzonych badań w ośrodkach zajmujących się badaniami reprezentacyjnymi w tym statystyką publiczną.

Możliwe kierunki dalszych badań, poszerzających tematykę poruszaną w ramach pracy obejmują takie zagadnienia jak zastosowanie innych niż regresyjne modeli uczenia maszynowego w statystyce małych obszarów, jak również zastosowanie dużych modeli językowych w kompleksowej realizacji badań próbkowych począwszy od projektowania kwestionariuszy ankiet, poprzez projektowanie badań i oczyszczanie danych, aż po wykorzystanie chmurowych rozwiązań AI do stawiania prognoz.