POLITECHNIKA WARSZAWSKA

WYDZIAŁ MATEMATYKI i NAUK INFORMACYJNYCH

ul. Koszykowa 75
00-662 WARSZAWA
NIP: 525-000-58-34
Tel.: (22) 234-79-88, 621-93-12
Fax.: (22) 625-74-60

Warszawa, 16 listopada 2021 roku

Prof. dr hab. inż. Jacek Mańdziuk
Wydział Matematyki i Nauk Informacyjnych
Politechnika Warszawska

Recenzja rozprawy doktorskiej mgr Barbary Żogały-Siudem zatytułowanej *Dobór zmiennych w modelach liniowych z wykorzystaniem indeksów wielowymiarowych*

Recenzja została przygotowana na prośbę Zastępcy Dyrektora Instytutu Badań Systemowych Polskiej Akademii Nauk, dr hab. inż. Jana W. Owsińskiego, wyrażoną w piśmie z dnia 20 września 2021 roku.

Tematyka rozprawy

Przedstawiona do recenzji rozprawa dotyczy metod selekcji zmiennych w modelach liniowych w kontekście problemów/danych o ekstremalnie dużej liczbie predyktorów. W takiej sytuacji zachodzi konieczność zastosowania metod efektywnego zarządzania danymi, w szczególności skutecznej pre-selekcji rozważanych zmiennych. Do tego celu Autorka stosuje indeksy wielowymiarowe umożliwiające szybkie wyszukiwanie wektorów sąsiadujących w rozważanej przestrzeni danych z danym wektorem wejściowym.

W rezultacie przeprowadzonych badań Doktorantka przedstawiła w rozprawie dwa algorytmy wykorzystujące indeksy wielowymiarowe – dedykowane sytuacjom, w których mamy do czynienia z ekstremalnie dużym wolumenem zmiennych - stanowiące rozwinięcie metod klasycznych: regresji krokowej oraz Lasso.

Poprawność formalna zaproponowanych algorytmów oraz praktyczne aspekty ich wykorzystania omawiane są w kolejnych rozdziałach rozprawy. W szczególności Autorka dowodzi, że zaproponowane rozszerzenia regresji krokowej oraz Lasso wykorzystujące indeksy wielowymiarowe zwracają wyniki dokładne w przypadku wykorzystania dokładnego indeksu wielowymiarowego. Ponadto, w oparciu o dane symulacyjne oraz rzeczywiste, Doktorantka pokazuje, że wykorzystanie indeksów wielowymiarowych w praktyce w kontekście ekstremalnie dużych zbiorów danych (kiedy nie są to indeksy dokładne) prowadzi do relatywnie niewielkiego spadku jakości budowanego modelu.

Hipoteza badawcza

Hipotezą badawczą weryfikowaną w rozprawie jest potwierdzenie możliwości efektywnego wykorzystania indeksów wielowymiarowych, w połączeniu z dwiema znanymi metodami doboru zmiennych do modelu predykcyjnego, w przypadku ekstremalnie dużych zbiorów zmiennych. Pojęcie „efektywnego wykorzystania" należy tutaj rozumieć jako uzyskanie modeli charakteryzujących się dobrą skalowalnością czasową i pamięciową kosztem zmniejszenia ich dokładności w akceptowalnym stopniu.

Treść rozprawy

Rozprawa liczy 136 stron, składa się z 6 rozdziałów oraz spisu literatury zawierającego 118 pozycji.

W pierwszym rozdziale Autorka przedstawia cel, zakres oraz motywację badań przedstawionych w rozprawie, wskazując jednocześnie te ich aspekty, które stanowią nowość w stosunku do aktualnego stanu badań w zakresie podjętego problemu.

Zagadnienie rozważane przez Kandydatkę jest bez wątpienia trudne i ma istotne znaczenie w kontekście wielu praktycznych zastosowań. Podejścia referencyjne, do których Autorka odnosi się w treści rozprawy są generalnie słabo skalowalne w kontekście rosnącej liczby predyktorów. Powyższa uwaga, w połączeniu ze spostrzeżeniem o lawinowo rosnącym wolumenie danych przechowywanych w różnego rodzaju repozytoriach prowadzi do natychmiastowego wniosku o potrzebie stosowania nowych podejść bądź odpowiednio zmodyfikowanych wersji istniejących metod. Powyższa konstatacja stanowi główną motywację praktyczną Autorki.

W dalszej części rozdziału pierwszego kandydatka formułuje podstawową hipotezę badawczą, opisuje metodykę prowadzenia przedstawionych w rozprawie badań oraz wprowadza jednolitą notację obowiązującą w całej dysertacji.

Rozdział napisany jest starannie. Jedyna uwaga dotyczy definicji średniej na stronie 13, w której pominięty został współczynnik 1/n.

W rozdziale drugim Doktorantka przedstawia podstawowe informacje dotyczące modeli liniowych, w tym dobór współczynników metodą najmniejszych kwadratów oraz odnosi się do podstawowego zagadnienia rozprawy, czyli selekcji zmiennych w modelu wielowymiarowym. Rozważania prowadzone są w trzech kontekstach: danych o małej liczbie zmiennych, danych o dużej liczbie zmiennych oraz danych o ultra-dużej liczbie zmiennych. Z punktu widzenia badań przestawionych w rozprawie najbardziej interesujący jest przypadek ostatni. Pojęcie ultra-dużej liczby zmiennych rozumiane jest jako sytuacja, w której liczba predyktorów rośnie szybciej niż wielomianowo z liczba obserwacji.

Rozdział ujmuje także omawiane w rozprawie zagadnienie w kontekście istniejącej literatury. Poza zapowiedzią modyfikacji algorytmów Lasso oraz metody regresji krokowej w przód pod kątem ich efektywnego wykorzystania w ekstremalnie dużych zbiorach zmiennych, rozdział zawiera informacje stosunkowo dobrze znane i mające charakter podstawowy. Ujęcie tematu jest właściwe, rozdział czyta się bardzo dobrze. Pominięte szczegółowe opisy metod można znaleźć w cytowanej literaturze. Jedyną wątpliwość wzbudziło u nie założenie o odwracalności macierzy X^TX prowadzące do wzoru (2.2), które nie zawsze musi być spełnione.

Rozdział trzeci poświęcony jest omówieniu indeksów wielowymiarowych, czyli struktur danych umożliwiających efektywne wyszukiwanie próbek znajdujących się w sąsiedztwie danego wektora (wektora zapytania) w rozważanej przestrzeni danych. Podstawowe dwa rodzaje zapytań dotyczą a) wskazania określonej z góry liczby najbliższych sąsiadów bądź b) wskazania wszystkich sąsiadów znajdujących się we wnętrzu kuli wielowymiarowej o zadanym promieniu, czyli tzw. zapytania zakresowego.

Autorka zaczyna od zwięzłego przedstawienia metod budowania indeksów dokładnych, skupiając się w dalszej części rozdziału na indeksach przybliżonych, ze swej natury bardziej predystynowanych do wykorzystania w ekstremalnie dużych zbiorach danych. Następnie Kandydatka omawia metody oparte o redukcję wymiarowości.

W wyniku przeprowadzonych poszukiwań optymalnej biblioteki implementującej metody indeksowania spełniające wymagania stawiane przez Autorkę (przede wszystkim możliwość definiowania pytań zakresowych oraz brak konieczności przechowywania całego zbioru danych w pamięci podręcznej) wybrana została biblioteka Faiss, której własności Autorka opisuje w dalszej części rozdziału.

Ostatnim elementem rozważań zamieszczonych w tej części pracy jest praktyczna analiza wpływu tzw. przekleństwa wymiarowości (*Curse of dimensionality*) w omawianym zagadnieniu. W kontekście przeprowadzonych przez Autorkę eksperymentów, zilustrowanych na rys. 3.7, rodzi się pytanie na ile reprezentatywne są zaprezentowane wykresy. Innymi słowy, czy wylosowanie kilku innych zestawów 50 zmiennych prowadziłoby, w sposób powtarzalny, do analogicznych wykresów, w szczególności chodzi o średnią zaznaczoną czarną linią na wykresie prawym? Jak wyglądałaby sytuacja w przypadku przeprowadzenia analizy dla wartości n innych niż 50? Uwaga czysto techniczna: dla wygody osoby czytającej warto byłoby umieścić na prawym rysunku także linię wykresu z rysunku lewego odpowiadającą n=50.

Kolejne dwa rozdziały zawierają najważniejsze wyniki pracy. Pierwszy z nich odnosi się do metody regresji krokowej w przód i zaczyna od szczegółowego omówienia oraz formalnego wprowadzenia metody w postaci Algorytmu 1 (w przypadku „standardowego" zastosowania metody, tj. w kontekście zbiorów zmiennych mniejszych niż rozważane w dysertacji zbiory ultra-duże). Następnie omawiane są popularne kryteria stopu algorytmu.

W dalszej części rozdziału przedstawione jest autorskie rozszerzenie rozważanej metody wykorzystujące indeks wielowymiarowy. Podstawowa idea polega na zawężeniu zbioru przeglądanych zmiennych w danym kroku metody do odpowiednio mniejszego podzbioru, który wyznaczany jest w oparciu o dokonaną wcześniej wielowymiarową indeksację danych.

Doktorantka formułuje i dowodzi prawdziwości czterech lematów, które następnie wykorzystuje w dowodzie sformułowanego przez siebie twierdzenia 4.5 odnoszącego się do wspomnianej wyżej modyfikacji Algorytmu 1. Zmodyfikowana wersja metody opisana jest w postaci Algorytmu 2. Następnie p. Żogała-Siudem pokazuje w Lemacie 4.6 równoważność Algorytmów 1 i 2 w przypadku kiedy mamy do czynienia z indeksem dokładnym.

Autorka omawia złożoność obliczeniową obu algorytmów, słusznie konkludując, że wymagania obliczeniowe Algorytmu 2 istotnie zależą od konkretnego zbioru danych i trudno tutaj o uniwersalne wnioski. Jednocześnie zwraca uwagę na fakt, że w przypadku, w którym dane nie

mieszczą się w pamięci roboczej, realizacja Algorytmu 1 będzie wymagała dokonywania częstych operacji wczytywania danych z dysku powodując dodatkowy narzut czasowy.

W dalszej części rozdziału przedstawiona jest analiza przydatności omawianych wcześniej kryteriów stopu w kontekście Algorytmu 2. Interesującym pomysłem jest opisanie tych kryteriów niejako w funkcji jednego bazowego kryterium nazwanego *kryterium w postaci spadku RSS*. Rozdział kończy listing Algorytmu 3 stanowiącego modyfikację Algorytmu 2 polegającą na efektywniejszym obliczaniu wartości progu η w kroku 7.

Niewątpliwie omawiany rozdział stanowi - obok rozdziału następnego - najistotniejszy fragment pracy i odnosi się do autorskich wyników p. Żogały-Siudem, sformułowanych w sposób ścisły (w postaci lematów i twierdzenia) i potwierdzonych formalnymi dowodami. Podobnie, Algorytmy 2 i 3 stanowią cenny wynik zarówno z teoretycznego jak i praktycznego punktu widzenia. Rozdział odnosi się bezpośrednio do postawionej tezy badawczej i udziela na zawarte w niej pytanie odpowiedzi twierdzącej na gruncie formalizmu matematycznego (wnioski praktyczne omówione są w rozdziale szóstym). Z drugiej strony Autorka przedstawia także swoje spostrzeżenia i intuicje, dzięki czemu rozdział nie jest zbyt hermetyczny i czyta się go z przyjemnością.

Z uwag polemicznych:

(1) Analiza *kryterium w postaci spadku RSS* mogłaby być pogłębiona, np. niektóre kryteria zależą wyłącznie od n, inne dodatkowo od p i/lub k. Czy powyższe różnice wpływają na praktyczną użyteczność tych kryteriów?

(2) W nierówności (4.7) pojawia się parametr η, którego znaczenie jest wyjaśniane dopiero podczas analizy Algorytmu 2.

Rozdział piąty omawia zagadnienie wyboru zmiennych w kontekście metody Lasso i algorytmu homotopijnego (Algorytm 4), który jest następnie modyfikowany przez Autorkę do postaci wykorzystującej indeksy wielowymiarowe (Algorytm 5). Motywacja leżąca u podstaw wykorzystania indeksów wielowymiarowych jest analogiczna do przypadku modyfikacji metody regresji krokowej – chodzi o ograniczenie liczby przeglądanych zmiennych (zdefiniowane albo wprost w postaci liczby sąsiadów albo w drodze ograniczeń nałożonych na odległość próbki od danego wektora).

W celu uzasadnienia poprawności zaproponowanej modyfikacji metody Lasso Kandydatka formułuje oraz dowodzi poprawności czterech twierdzeń pomocniczych (lematy 5.2-5.5), które wykorzystuje w dowodzie twierdzenia 5.6 odnoszącego się wprost do właściwości procedury wyboru kolejnej zmiennej.

W dalszej części rozdziału Autorka bada złożoność zaproponowanej metody budowy modelu, czyli Algorytmu 5, bada wpływ niedokładności indeksu na optymalność wyboru ostatecznego zestawu zmiennych oraz łączy zaproponowaną metodę z algorytmem przesiewowym *Screening-Ordering-Selection* (SOS) w celu umożliwienia wyboru mniejszego zestawu zmiennych (prostszego modelu).

W ostatniej części rozdziału Doktorantka opisuje alternatywne metody przesiewania zmiennych stosowane w algorytmie Lasso oraz dokonuje ich porównania z metodą SOS (w kontekście Algorytmu 5).

Jak wspomniałem powyżej rozdział piąty jest kluczowym, obok rozdziału czwartego, fragmentem dysertacji. Doktorantka przedstawia w nim autorski algorytm wykorzystania indeksu wielowymiarowego w popularnej metodzie doboru zmiennych Lasso, umożliwiający zastosowanie tej

metody do problemów o znacznie większych rozmiarach. Zarówno algorytm jak i jego własności opisane są w sposób formalny po raz kolejny pokazując bardzo dobre rozumienie przez Kandydatkę rozważanej tematyki oraz swobodę posługiwania się formalizmem matematycznym. Ponownie warto podkreślić, że proporcje pomiędzy opisem formalnym a opisem intuicyjnym są dobrze dobrane.

Mam dwie uwagi do treści rozdziału piątego:

(1) Na początku rozdziału (str. 67) rozpatrywane są dwa równania: (5.1) oraz równanie poniżej (bez numeru). Równania te określa Autorka jako „równoważne". Oczywiście, optymalizacja w obu przypadkach prowadzi do tego samego rezultatu, ale z formalnego punktu widzenia nie są to zapisy równoważne. Tym samym, w równaniu dolnym nie powinno być stosowane oznaczenie β_{reg} ale inne oznaczające optymalizowaną funkcję.

(2) Z podpisu pod rysunkiem 5.3 wynika, ze „lewy wykres pokazuje rozwiązania mniejsze, prawy większe", natomiast w tekście zamieszczonym kilka linii powyżej pada stwierdzenie odwrotne. Nota bene, rysunki lewy i prawy są w zasadzie nierozróżnialne, a w konsekwencji trudno wyciągnąć jakiekolwiek wnioski z ich porównania.

Kolejny rozdział, w odróżnieniu od poprzednich, które poświęcone były wprowadzeniu oraz teoretycznemu uzasadnieniu autorskich algorytmów wykorzystujących indeksy wielowymiarowe, poświęcony jest praktycznej weryfikacji skuteczności proponowanych rozwiązań, w oparciu o bazę Eurostatu zawierającą ok. 8 milionów zmiennych.

Analiza eksperymentalna zaczyna się od testów na danych symulacyjnych, które odnoszą się do faktycznych kolumn zmiennych Eurostatu, dla których zmienne odpowiedzi generowane są sztucznie w oparciu o 3 modele (A, B, C). Modele te implementują kombinację liniową odpowiednio dwóch, trzech i czterech zmiennych z dodanym szumem gaussowskim.

W przypadku regresji krokowej (w wersji zmodyfikowanej przez Autorkę) badana jest skuteczność algorytmu przesiewania zmiennych, jakość predykcji uzyskanego modelu oraz wpływ kryterium stopu na funkcjonowanie modelu i jakość uzyskiwanych wyników.

Zmodyfikowany algorytm Lasso również testowany jest pod kątem skuteczności przesiewania zmiennych i jakości predykcji. Zastosowany algorytm przesiewania jest porównywany z innymi metodami przesiewania wykorzystywanymi w metodzie Lasso.

W dalszej części rozdziału Autorka porównuje między sobą oba algorytmy w kontekście zestawu wyselekcjonowanych zmiennych, czasu wykonania oraz jakości predykcji.

Rozdział kończy opis przypadku rzeczywistego czyli wykorzystania obu algorytmów do budowy modeli dla konkretnej zmiennej w oparciu o bazę Eurostatu. Doktorantka analizuje zestawy zmiennych zaproponowane przez oba modele, a także odnosi się do – istotnego z praktycznego punktu widzenia – zagadnienia tzw. *przecieków informacyjnych*.

Podsumowując tę część pracy musze stwierdzić, że rozdział szósty stanowi pewne rozczarowanie. Spodziewałem się, że Kandydatka dokona szerszej analizy efektywności zaproponowanych przez siebie algorytmów w oparciu o dane rzeczywiste. Zamiast tego znakomita większość rozdziału poświęcona jest testom na danych symulacyjnych (jedynie jeden test w pełni dotyczy rzeczywistych danych Eurostatu). Rozumiem powody takiej decyzji związane zapewne z uciążliwością prowadzenia eksperymentów na danych rzeczywistych (kwestia dostępu, korelacji zmiennych, braków i szumów w danych, itd.) niemniej jednak, jedynie takie testy dają odpowiedź na pytanie o *praktyczną wartość* zaproponowanych przez Doktorantkę metod.

Uwagi szczegółowe:

(1) Autorka kilkukrotnie wspomina w pracy o tym, że koszt zaindeksowania bazy jest jednorazowy i w związku z tym nie jest uwzględniany w analizie złożoności algorytmów, z czym można się zgodzić, niemniej dla kompletności opisu warto byłoby podać przybliżony koszt operacji indeksowania bazy Eurostatu.

(2) Doktorantka przyjmuje założenie o małej zmienności w czasie bazy Eurostatu, dzięki czemu możliwe jest skorzystanie z jednokrotnego indeksowania bazy. Na ile to założenie jest w praktyce spełnione? Wydaje się, że jest ono silnie zależne od wyboru konkretnej opisywanej zmiennej.

(3) Na stronie 97 Autorka pisze, że „Braki na wykresie dla modeli symulacyjnych B i C wynikają z kryterium stopu ..." nie podając jakie kryterium było zastosowane.

(4) Na stronie 103 (przedostatnia linia) Autorka nie wyjaśnia dlaczego przyjmuje jako wartość graniczną 0.3.

(5) Na stronie 108 jako wartość graniczną bezwzględnej korelacji z faktycznie użytą zmienną Kandydatka przyjmuje 0.99. Jak zmieniłyby się wyniki gdyby przyjąć inne wartości tego progu, np. 0.98 czy 0.995? Generalnie, wartość ta istotnie zależy od konkretnego zbioru danych. W jaki sposób powinna być dobierana w ogólnym przypadku?

(6) Na stronie 110, w opisie rysunku 6.14 punkty poniżej przekątnej odpowiadają wyborowi zmiennej x_0 w oparciu o r_L. Autorka wskazuje, że jest to 56% przypadków. Patrząc na rysunek wydaje się, że liczba ta jest większa.

(7) Na stronie 118 Doktorantka porusza bardzo ważny problem tzw. przecieków informacyjnych, wskazując, że w badanym przypadku pominięte zostały określone dane z bazy Eurostatu z uwagi na istnienie wspomnianego zjawiska. Baza Eurostatu jest ekstremalnie duża i posiada stosunkowo rozbudowaną strukturę. Z czego wynika przekonanie Doktorantki, że wszystkie zmienne, które w tym kontekście należało pominąć zostały faktycznie pominięte?

(8) Ostatnią kwestią, na którą chciałbym zwrócić uwagę w tym rozdziale jest istotna różnica pomiędzy wynikami uzyskanymi przy wykorzystaniu obu metod: MI-Lasso oraz MI-ForwardStepwise. W przypadku zapytania dotyczącego śmiertelności noworodków metoda pierwsza wybiera zmienne dość jednorodne (z obszarów ekonomii oraz zmiennych populacyjnych) dzięki czemu dokonany wybór jest bardziej intuicyjny. Z kolei druga metoda wybiera bardzo zróżnicowane zmienne, przy czym niektóre z nich nie mają oczywistego związku ze zmienną opisywaną (np. *udział pasażerów podróżujących pociągami* czy *agregaty rachunków narodowych dla innych, niesklasyfikowanych inaczej działalności usługowych*). Powyższe różnice wydają się być kluczowe z punktu widzenia oceny obu metod, np. wyjaśnialności modelu. Jeden przykład testowy nie pozwala niestety na żadne uogólnienia czy wyciągnięcie ugruntowanych wniosków.

Rozdział siódmy stanowi podsumowanie dysertacji. Autorka przypomina motywację leżącą u podstaw prowadzonych badań oraz streszcza najistotniejsze wyniki. Pewnym mankamentem jest brak odniesienia do możliwych kierunków dalszych badań w rozważanym przez Autorkę temacie. Rozprawę dopełnia spis literatury.

Rozprawa napisana jest starannie, warstwa językowa jest na wysokim poziomie. Nieliczne błędy językowe czy interpunkcyjne, które zauważyłem zdecydowanie mieszczą się w dopuszczalnym zakresie.

Oryginalny wkład Autorki rozprawy

Oryginalny wynik badawczy Doktorantki dotyczy sformułowania, analizy własności, implementacji oraz eksperymentalnej weryfikacji rozszerzeń dwóch klasycznych metod budowy modeli predykcyjnych bazujących na wykorzystaniu indeksów wielowymiarowych. W efekcie, rozszerzone wersje obu metod mogą być skutecznie zastosowane w przypadku danych o ultra-dużej liczbie zmiennych, które z uwagi na swój rozmiar są poza zasięgiem wersji bazowych tychże metod.

Przedstawione w rozprawie wyniki badawcze zostały częściowo opisane w czterech publikacjach (trzy prace są opublikowane, jedna jest w recenzji).

Konkluzja

Przedstawiona do recenzji rozprawa zawiera oryginalne wyniki prac badawczych p. Barbary Żogały-Siudem w obszarze doboru zmiennych w modelach liniowych w przypadku ultra-dużych zbiorów zmiennych. Nie dostrzegłem w rozprawie istotnych nieprawidłowości, a wymienione w recenzji uwagi nie wpływają na moją ogólnie wysoką ocenę dysertacji.

Rozprawa dotyczy aktualnej tematyki badawczej a jej treść bez wątpienia dowodzi wiedzy Autorki w zakresie rozważanych zagadnień. Tematyka oraz treść rozprawy mieszczą się w obszarze dyscypliny *informatyka techniczna i telekomunikacja.*

Reasumując, **stwierdzam, że rozprawa spełnia wymagania stawiane przez odnośną Ustawę i wnoszę o jej przyjęcie oraz dopuszczenie jej Autorki, mgr inż. Barbary Żogały-Siudem, do dalszych etapów przewodu doktorskiego.**