Dr hab. inż. Wojciech Kotłowski, prof. PP
Instytut Informatyki (Institute of Computing Science)
Politechnika Poznańska (Poznań University of Technology)
ul. Piotrowo 2, 60-965 Poznań
tel: (+48) 61 665 2936
wkotlowski@cs.put.poznan.pl

Poznań, 31 stycznia 2022 roku

Recenzja rozprawy doktorskiej

mgr Barbary Żogały-Siudem

Dobór zmiennych w modelach liniowych z wykorzystaniem indeksów wielowymiarowych
(Variable selection algorithms for linear models based on multidimensional indices)

1 Wprowadzenie

Rozprawa doktorska mgr Barbary Żogały-Siudem dotyczy problemu selekcji zmiennych w modelach liniowych budowanych na bardzo dużej („ekstremalnej") liczbie predyktorów. Przez ekstremalną liczbę zmiennych Autorka rozumie sytuację, w której ich liczba p nie tylko przewyższa liczbę obserwacji n w zbiorze danych, ale przewyższa ją o wiele rzędów wielkości, np. w bazie danych Eurostatu można odnaleźć 8 milionów zmiennych, przyporządkowanym 800 obserwacjom. Autorka wskazuje też, że z punktu widzenia teoretycznego przypadek ultra-wymiarowości zbioru danych zachodzi, gdy p skaluje się wykładniczo względem n.

Już sam fakt przewyższania liczby obserwacji przez liczbę predyktorów (tzw. przypadek *large p, small n*) powoduje pojawianie się trudności natury statystycznej i obliczeniowej. Wiele metod może w tym przypadku po prostu zawieść – np. w regresji liniowej napotykamy na problem z odwracalnością macierzy Grama, w regresji logistycznej dane są trywialnie separowalne liniowo, przez co wektor współczynników rozbiega się do nieskończoności, itp. Z teoretycznego punktu widzenia, liczba parametrów (np. wag w modelu liniowym) przekraczająca liczbę obserwacji utrudnia lub uniemożliwia podanie gwarancji na jakość modelu (np. istotność statystyczna, trafność predykcji poza zbiorem uczącym) bez poczynienia dodatkowych założeń o rozkładzie danych. Spowodowało to rozwój metod dedykowanych do rozwiązywania tego typu problemów, które w ogólności bazują na wyborze stosunkowo niewielkiego podzbioru zmiennych i zbudowania modelu predykcyjnego tylko na podstawie takiego podzbioru. Metody takie jako uzasadnienie swojego działania zakładają, że zmienna objaśniana jest funkcją kombinacji niewielkiej liczby predyktorów, a pozostałe z predyktorów są nieinformatywne. Najbardziej znaną metodą tego typu jest zapewne regresja Lasso, która wraz z minimalizacją sumy kwadratów błędów nakłada karę

na współczynniki regresji w postaci sumy ich wartości bezwzględnych, wymuszając w ten sposób zerowanie się sporej ich części. Innym przykładem są metody regresji sukcesywnie dobierające kolejne predyktory do modelu z określonym warunkiem stopu.

W rozważanym przez p. Żogałę-Siudem przypadku mamy jednak do czynienia z sytuacją ekstremalną, w której liczba zmiennych jest na tyle duża, że samo ich przeglądanie zaczyna być skomplikowane pod względem obliczeniowym. W związku z tym praca skupia się na przyspieszeniu działania algorytmów dedykowanych do problemów $p \gg n$ tak, aby były w stanie radzić sobie z milionami cech i wybrać spośród nich niewielki podzbiór kilku do kilkudziesięciu cech istotnych. W szczególności, algorytmy powinny – z uwagi na rozmiar danych – działać bez konieczności ładowania pełnego zbioru danych do pamięci operacyjnej i pozyskiwać wartości danego predyktora na żądanie z pamięci zewnętrznej tylko w sytuacji, gdy będzie on potencjalnie potrzebne. Odbywa się to poprzez zastosowanie tzw. *indeksów wielowymiarowych*, struktur danych, których celem jest zwrócenie predyktorów „podobnych" do pożądanego wektora (np. wektora residuów aktualnego modelu) bez konieczności liczenia na bieżąco podobieństwa i przeglądania wszystkich możliwych cech.

W pracy, dzięki bardzo trafnie użytej operacji normalizacji wektorów predyktorów, najsensowniejszą miarą „podobieństwa" okazuje się zastosowanie iloczynu skalarnego wektorów, który jest przy normalizacji równoważny współczynnikowi korelacji. Co więcej, maksymalizacja/minimalizacja iloczynu skalarnego okazuje się równoważna do znajdywania wektorów najbliższych w sensie miary odległości euklidesowej. To powoduje, że Autorka może korzystać z obszernej literatury i wielu rozwiązań algorytmicznych dotyczących tworzenia indeksów do szybkiego znajdywania najbliższych sąsiadów, zarówno w wersji dokładnej jak i przybliżonej.

Oczywiście, sama konstrukcja takiego indeksu wielowymiarowego jest dodatkowym narzutem i wymaga uprzedniego policzenia odległości między wektorami, a następnie utworzenia specjalnej struktury danych (np. drzewa bądź wektorowych sygnatur w tablicy haszowej). Autorka pracy dobrze uzasadnia jednak taki narzut, ponieważ czas na budowę indeksu musi zostać poświęcony tylko raz, a później można z niego stale korzystać, budując wiele interesujących modeli. Celem jest więc zautomatyzowanie procesu modelowania zbioru danych o bardzo dużej liczbie zmiennych z zastosowaniem indeksów wielowymiarowych. W mojej opinii cel ten jest bardzo dobrze uzasadniony, interesujący, o dużym wymiarze praktycznym, a przede wszystkim wystarczająco ambitny, aby stał się przedmiotem rozprawy doktorskiej.

2 Ocena struktury i zawartości pracy

Recenzowana rozprawa jest napisana bardzo czytelnie. Mimo sporej liczby wprowadzanych pojęć i wielu wyników formalnych (definicji, dowodów, twierdzeń), oraz bardzo obszernej analizy eksperymentalnej, czyta się ją bardzo płynnie i jest również wyjątkowo staranna pod względem języka, a wszystkie wykresy i rysunki są wykonane w sposób wzorowy (sam chciałbym umieć w ten sposób wizualizować wyniki analizy danych). Praca liczy 136 stron i składa się z siedmiu rozdziałów. Całościowo jej strukturę oceniam całkowicie pozytywnie, a sam podział na rozdziały wydaje się zupełnie naturalny i zgodny z podziałem tematycznym treści. Poniżej omówię zawartość poszczególnych rozdziałów, równocześnie starając się dokonać ich merytorycznej oceny.

Rozdział pierwszy ma charakter wprowadzenia. Opisany w nim zostały cel i motywacja pracy, a także określone elementy nowości w rozprawie względem stanu zastanego. W szczególności p. Żogała-Siudem opisuje istniejące narzędzia do wyszukiwania i selekcji atrybutów, usługę *Google Correlate*, będącą inspiracją do prowadzonych badań, a także

klasyczne już metody selekcji cech w modelach liniowych, tzw. regresję krokową w przód oraz regresję Lasso. Sformułowana jest również wspomniana we wprowadzeniu hipoteza badawcza (wykorzystanie przybliżonych indeksów wielowymiarowych w metodach doboru zmiennych do modelu znacznie je przyspiesza kosztem bardzo nieznacznego spadku na ich jakości). Metodologia badawcza obejmuje zarówno analizę teoretyczną, jak i obszerne eksperymenty obliczeniowe. Pojawia się również krótka sekcja opisująca układ pracy oraz zastosowaną w pracy notację. Układ rozdziału przyjmuję bez zastrzeżeń, a pomysł wprowadzenia na wstępie notacji – później jednolicie i całkowicie spójnie stosowanej w całym tekście – jako bardzo ułatwiający czytanie.

Rozdział drugi wprowadza czytelnika w *state-of-the-art* dziedziny, tzn. opisuje modele liniowe wraz z istniejącymi najbardziej popularnymi metodami selekcji zmiennych. Rozpoczyna się zwięzłym opisem problematyki regresji liniowej, wraz z metodą najmniejszych kwadratów, oraz geometryczną interpretacją rozwiązania jako rzutowania wektora zmiennej objaśnianej na podprzestrzeń rozpiętą przez wektory wejściowe. W dalszej części opisane są miary oceny jakości modelu liniowego w postaci współczynnika determinacji, statystyki F (i stowarzyszonego z nią testu statystycznego), a także kryterium dodawania zmiennych w modelu regresji krokowej w postaci (relatywnej) zmiany wartości sumy kwadratów błędów (*RSS*). Następnie omawiane są metody selekcji zmiennych – zarówno dla przypadku małej liczby zmiennych ($p < n$), jak i przypadkach dużej i wreszcie ekstremalnej liczby zmiennych $p \gg n$. W pierwszym przypadku istnieje możliwość bezpośredniego testowania wszystkich możliwych podzbiorów, bądź użycie regresji krokowej sukcesywnie dodając (bądź odejmując) po jednej zmiennej, zaczynają z pustego (bądź pełnego) zbioru zmiennych, każdorazowo przeliczając model liniowy i wyznaczając wartość RSS dla każdego kandydata. W drugim przypadku metody intensywne obliczeniowo zawodzą i stosuje się zwykle metody prostsze: *Orthogonal Matching Pursuit (OMP)*, która wybiera w każdym kroku zmienną najbardziej skorelowaną z aktualnym wektorem reszt, następnie uaktualnia model; *Forward Stagewise Regression*, która, dodaje do modelu nową zmienną ze współczynnikiem wyliczonym wyłącznie na podstawie jej korelacji z aktualnym wektorem reszt, bez patrzenia na żadne inne współczynniki; wreszcie metodę Lasso, która wprowadza dodatkowy człon regularyzacyjny w postaci suma wartości bezwzględnych do kryterium optymalizacji, dzięki czemu model może być dopasowywany do danych przy pełnej liczbie zmiennych a równocześnie wymusza na wektorze współczynników postać rzadką, tzn. tylko niektóre współczynniki modelu są różne od zera. W dalszej kolejności opisywane są nieliczne metody działające przy ultra-dużej liczbie zmiennych: *Sure Independence Screening (SIP)*, polegająca na filtrowaniu zmiennych na podstawie korelacji z wektorem odpowiedzi i wybraniu do modelu podzbioru zmiennych o ustalonej liczności; regresja krokowa w przód jako metoda filtrowania nieistotnych zmiennych; metody wstępnego przesiewania zmiennych do rozwiązania problemu Lasso; wreszcie strumieniowe przetwarzanie zmiennych. Na koniec Autorka motywuje swoje własne wyniki badawcze nawiązując do przypadku, w którym wielokrotnie wykorzystuje się te same predyktory (np. wykorzystując je do budowy dużej liczby modeli dla różnych zmiennych odpowiedzi) i braku satysfakcjonujących rozwiązań tego przypadku w literaturze. Rozdział nie opisuje nowych wyników, ale świetnie wprowadza do dalszej części pracy, przy okazji wskazując, że Autorka ma obszerną wiedzę w zakresie dziedziny selekcji i filtrowania zmiennych w modelach liniowych.

W rozdziale trzecim opisane są indeksy wielowymiarowe. Wychodząc z idei szybkiego wyszukiwania najbliższych sąsiadów względem metryki euklidesowej, Doktorantka wskazuje, że dla wektorów znormalizowanych kryterium to jest równoważne maksymalizowaniu korelacji (a wykonując je również dla zanegowanego wektora – również minimalizacji korelacji). Następnie omawiane są podstawowe rodzaje indeksów z podziałem na dokładne i przybliżone. Reprezentantami tych pierwszy są drzewa KD, a także podobnie zmotywowane drzewa kulowe czy drzewa metryczne. Autorka bardzo rozsądnie uzasadnia jednak

rezygnacje z indeksów dokładnych z powodu dużego wymiaru rozważanego problemu selekcji zmiennych (wymiarem jest tu akurat liczba obserwacji n). Stąd dokładniej omówione zostają indeksy przybliżone: *Locality Sensitive Hashing* polegające na przyporządkowaniu punktów do kubełków przy użyciu funkcji haszującej respektującej wzajemne podobieństwo punktów; kwantyzację produktową polegającą na podziale współrzędnych wektora na grupy i kodowanie każdej z grup oddzielnie; grafy sąsiedztwa, w którym punkty są wierzchołkami, a krawędzie łączą wierzchołki z jego najbliższymi sąsiadami; struktury drzewiaste działające podobnie do drzew KD, ale zwracające rozwiązania przybliżone. Następnie opisane są metody oparte o redukcję wymiarowości, takie jak losowe rzutowanie do przestrzeni o mniejszym wymiarze, czy analiza składowych głównych. Najwięcej uwagi Autorka poświęca jednak bibliotece Faiss, implementujące tzw. indeksy odwrócone. Polegają one na grupowaniu wektorów w klastry i przechowywaniu centroidów klastrów w osobnym indeksie przestrzennym, dzięki czemu przy znajdywaniu najbliższego sąsiada wpierw znajduje się stosunkowo niewielką liczbę najbliżej leżących centroidów, a później dla każdego z centroidów można przeszukać jego klaster w poszukiwaniu wektorów najbardziej podobnych.

Rozdział zawiera również ciekawą pracę własną Doktorantki poświęconą eksperymentalnej analizie zapytań za pomocą biblioteki Faiss, z wykorzystaniem zmiennych z bazy danych Eurostat. W ramach eksperymentu badane są odsetek poprawnie znalezionych zmiennych w stosunku do wszystkich, które należało znaleźć, w funkcji minimalnej zadanej korelacji, a także parametru *nproble* algorytmu okeślającego liczbę przeglądanych klastrów. Równocześnie badany jest również – w funkcji tych samych zmiennych – odsetek zwróconych zmiennych (spośród wszystkich zmiennych), także sam czas wykonywania zapytania względem *nprobe* oraz liczby zwróconych zmiennych. Dzięki tym eksperymentom, Doktorantka jest w stanie określić sensowne parametry algorytmu na potrzeby testów w dalszych etapach pracy. Rozdział kończy się interesującą analizą tzw. „klątwy wymiarowości" czyli faktu, że wraz z rosnącym wymiarem punkty znajdują się coraz dalej od siebie. Generując wektory z rozkładu jednostajnego na sferze, Autorka wyznacza teoretycznie procent punktów leżących w odległości (mierzonej za pomocą korelacji) nie dalszej niż zadany próg, a następnie wykreśla ten procent w funkcji progu i wymiaru. Wyniki te porównane są z danymi z Eurostatu i pokazują istotne różnice, sugerujące, że dane z Eurostatu mają naturalne skupienia i autokorelacje. Na koniec pojawia się również krótka analiza przypadkowych korelacji, tzw. korelacje pojedyncze i wielokrotne, które występują między niezależnymi zmiennymi losowymi na skutek dużej liczności zbioru tych zmiennych. Eksperymenty te uważam za ciekawe i warte osobnego opublikowania, ponieważ pokazują znaczące różnice rozkładu znormalizowanych wektorów z rzeczywistego zbioru danych względem zbioru wektorów losowych, a także zwracają uwagę na siłę zjawiska występowania dużych, przypadkowych korelacji w licznym zbiorze wektorów.

Rozdział czwarty dotyczy zastosowań indeksów wielowymiarowych do przyspieszenia metody regresji krokowych i stanowi w dużej mierze pracę własną Autorki. We wstępie Doktorantka opisują podstawowy algorytm regresji krokowej, który rozpoczyna od pustego zbioru zmiennych, a następnie w każdej iteracji dodaje jedną nową zmienną, która maksymalizuje spadek kryterium RSS. Dodatkowo, algorytm musi zostać wyposażony w kryterium stopu (ponieważ RSS zawsze spada). Opisana została tutaj klasa kryteriów w postaci karanego logarytmu wiarogodności modelu, gdzie konkretną postać kryterium dla problemu regresji liniowej determinuje wybór funkcji kary. Wymienione i opisane zostały najpopularniejsze funkcje kary takie jak kryteria Schwarza (*BIC* oraz modyfikacja w postaci *extended BIC*, Akaikego (również z modyfikacją), Hannana-Quinna i Mallowsa. Wybiegając naprzód, pod koniec rozdziału Autorka dodaje również kryterium oparte na teście F (z bardzo rozsądnie skorygowanym poziomem istotności na testy wielokrotne w oparciu o poprawkę Bonferroniego), a także kryterium oparte na skorygowanym współczynniku determinacji. Następnie wszystkie te kryteria w sposób jednolity opisywane są jako określające próg na

względnym spadku RSS, gdzie sama postać progu jest pewną, zależną od kryterium, funkcją liczby obserwacji n, aktualnej liczby predyktorów k czy całkowitej liczby predyktorów p. Jest to bardzo czytelne zestawienie różnorodnych kryteriów stopu, pozwalających dowolne z nich wykorzystać w sposób „automatyczny" w algorytmie regresji krokowej.

W rozdziale Autorka przeprowadza teoretyczną analizę obliczeń wykonywanych przez algorytm regresji krokowej, pozwalającą sprowadzić jego działanie do badania wartości korelacji. Wpierw, w lemacie 4.1 przepisane zostaje wyrażenie na różnice sumy kwadratów błędów przy dodaniu nowego wektora x jako korelacja wektora odpowiedzi y z unormowanym rzutem wektora x na podprzestrzeń rozpiętą przez wektory już dodane, co przy wykorzystaniu ortogonalności tych wektorów daje się z kolei przepisać (lemat 4.2) jako iloraz wyrażeń zawierających korelacje wektora x z rzutem wektora odpowiedzi i wektorami z modelu. Pozwala to z kolei w lemacie 4.3 na ustalenie warunku koniecznego na korelacje tworzone przez wektor x z wektorami z modelu przy założonym względnym spadku różnicy sumy kwadratów błędów (lemat 4.4 pokazuje, że warunku tego nie można w ogólności polepszyć). Stanowi to bazę do przesiewania zmiennych w algorytmie regresji krokowej, ponieważ założony względny spadek różnicy sumy kwadratów błędów można uzyskać:

- wybierając pewną sensowną zmienną kandydata x_0 i ustalając względny spadek RSS po jej dodaniu (twierdzenie 4.5), lub
- ograniczając od dołu względny spadek korzystając z jednego z kryteriów stopu (twierdzenie 4.7) i wspomnianego już jednolitego zapisania wszystkich kryteriów jako progów na względnym spadku RSS, lub
- korzystając z obydwu ograniczeń na raz (biorąc lepsze z nich).

Uważam powyższe pomysły za bardzo ciekawe. Co prawda, ograniczenie poprzez wybór kandydata jest w ogólności tak dobre jak wybrany kandydat, to jednak w eksperymentach obliczeniowych okazuje się, że wybór kandydata prostymi heurystykami, takimi jak korelowanie się kandydata z wektorem reszt modelu, okazuje się często bardzo dobre, albo wręcz trudne do poprawy. Korzystając z wyprowadzonego warunku koniecznego i ograniczenia na względny spadek RSS Autorka uzyskuje efektywnie bardzo dobrą metodę odsiewania zmiennych wyłącznie w oparciu o korelacje z wektorami z danych, co w parze z użyciem szybkich indeksów wielowymiarowych pozwala znacznie przyspieszyć algorytm regresji krokowej: zamiast milionów zmiennych wymagających testowania w każdym kroku (co każdorazowo wiąże się, zgodnie z lematem 4.2, z policzeniem korelacji zmiennej z rzutem r wektora y oraz z wszystkimi ortogonalizowanymi zmiennymi obecnymi już w modelu), liczba zmiennych-kandydatów lepszych od x_0 może zostać łatwo zmniejszone o rzędy wielkości. Finalnym efektem rozdziału są algorytmy MI-ForwardStepwise (oparty na wyborze kandydata x_0) oraz MIC-ForwardStepwise (oparty dodatkowym uwzględnieniu kryterium stopu przy odsiewaniu). Ich działanie oparte jest szczegółowej analizie teoretycznej, pozwalającej wyznaczyć kryteria jakimi posługuje się oryginalny algorytm wyłącznie poprzez korelacje między nową i starymi zmiennymi, a następnie przy użyciu wyprowadzonych warunków odsiewania zmiennych użyć całej maszynerii indeksów odwrotnych. Co więcej, formalnie pokazano, że jeśli indeksy wielowymiarowe byłyby dokładne, to powyższe algorytmy działają *identycznie* do oryginalnego algorytmu regresji krokowej (lemat 4.6), a więc niedokładność indeksu jest jedynym miejscem, które potencjalnie zaburza działanie algorytmu.

Rozdział zawiera również porównanie złożoności obliczeniowej zaproponowanych algorytmów względem parametrów problemu (n, p, k) oraz wielkości zakresu punktów zwracanych każdorazowo przez zapytanie do indeksu, a także porównanie przebiegu poszczególnych kryteriów stopu względem liczności próbki n czy liczby dodanych do modelu zmiennych k. O rozdziale i wynikach w nim zawartych mam zdanie bardzo pozytywne: doceniam

wnikliwy wgląd w algorytm regresji krokowej z przeformułowaniem kryteriów (spadek RSS, progi implikowane przez reguły stopu) do postaci korelacji i pomysłowego ich zastosowania do użycia indeksów wielowymiarowych.

Rozdział piąty poświęcony jest zastosowaniu indeksów wielowymiarowych do algorytmu Lasso. Rozpoczyna się od zdefiniowania kryterium optymalizacji Lasso i omówienia ogólniejszej metody regularyzacji za pomocą norm ℓ_q, poprawnie wskazując, że norma $q = 1$ jest jedyną normą indukującą rzadkość wektora współczynników, która równocześnie prowadzi do wypukłego problemu optymalizacji. Następnie przedstawiony zostaje algorytm homotopijny, który pozwala nie tylko na rozwiązanie problemu Lasso z ustaloną siłą regularyzacji λ, ale zwraca całą ścieżkę regularyzacji, czyli przebieg wartości wag parametrów gdy λ zmniejszane jest od nieskończoności (efektywnie: najmniejszej wartości λ dla której wagi wszystkich współczynników w modelu są zerowe) do zera. Jest to możliwe, ponieważ ścieżka regularyzacji jest odcinkami liniowa, co jest w sposób czytelny wyprowadzone przez Autorkę z warunków optymalności. Następnie przedstawione zostało wyznaczenie ścieżki regularyzacyjnej przy użyciu indeksów wielowymiarowych, przy czym metodzie przyświeca ta sama zasada, co w rozdziale poprzednim: zakładając, że indeks jest dokładny, nowy algorytm powinien działać identycznie z oryginalną metodą Lasso. Lemat 5.2 pokazuje, że wartość λ, przy której następuje dodanie nowej nieaktywnej zmiennej może zostać zapisana jako równanie na korelacje nowej zmiennej z dwoma zmiennymi utworzonymi ze zmiennych znajdujących się już w zbiorze danych. Lemat 5.3 daje ograniczenia na współczynniki w równaniu korelacji. Dodatkowo, lemat 5.4 pozwala wyznaczyć warunek konieczny dla wszystkich zmiennych nieaktywnych ze względu na wartość parametru regularyzacji, przy której nastąpił ostatni skok. W końcu lemat 5.5 mówi, jak dowolna zmienna nieaktywna x_0 i stowarzyszona z nią wartość λ_0, przy której byłaby wprowadzona do modelu, dają dodatkowe ograniczenia na wartości λ przy których mogłaby być wprowadzona do modelu inna zmienna nieaktywna. Cechą wspólną wszystkich tych ograniczeń jest to, że wyrażone są poprzez wartości korelacji wektorów zbudowanych na podstawie zmiennych z modelu i zmiennej-kandydata x_0, a więc pozwalają użyć do przesiewania zmiennych indeksu wielowymiarowego, co jest treścią twierdzenia 5.6 i bazą do nowej wersji algorytmu homotopijnego z zastosowanie indeksów wielowymiarowych.

Co ciekawe, w tym rozdziale użyte zostały podobne triki jak w rozdziale poprzednim: wyrażenie kryteriów używanych przez algorytm w postaci korelacji, wyznaczenie ograniczeń na te korelacje uzyskanych przez wprowadzenia rozsądnej zmiennej-kandydata, a następnie przesiewanie zmiennych nieaktywnych aby otrzymać tylko takie, które byłyby wprowadzone przed zmienną-kandydatem, używając indeksu wielowymiarowego. Pokazuje to tylko, że dobre pomysły znajdują zastosowanie wielokrotnie. Podobnie jak poprzednio, analiza działania algorytmu Lasso i dowody warunków koniecznych na zmienne nieaktywne formalnie pokazują, że nowa, bardziej efektywna wersja metody, działa identycznie do oryginału przy założeniu dokładności indeksu. Rozdział zawiera również analizę złożoności algorytmu, rozwiązanie problemów z niedokładnością indeksu (tym razem potencjalnie bardziej niebezpieczną, ponieważ powoduje zejście ze ścieżki regularyzacjnej), a także przegląd metod tzw. wstępnego przesiewania zmiennych, pozwalających odrzucić zmienne, które nie będą aktywne w rozwiązaniu.

Rozdział szósty poświęcony jest analizie eksperymentalnej zaproponowanych algorytmów. Opiera się on a bazie Eurostatu, już wcześniej wspominanej i używanej w pracy, z której uzyskane zostaje 800 obserwacji (szeregi czasowe dotyczące państw europejskich) oraz 8 milionów cech. Wpierw Autorka opisuje problemy związane z silnymi korelacjami między zmiennymi, między innymi obecność cech-duplikatów, cech wyrażonych w różnych jednostkach, kopii zmiennej wyjściowej powodujące przeciek informacji, itp. Konkluzją jest konieczność pewnej dozy interakcji systemu z użytkownikiem, ponieważ problemów powyższych nie da się wyeliminować automatycznie (tj. wyłącznie na podstawie samych da-

nych). Opisane jest również przygotowanie danych do analizy, m.in. wypełnianie wartości brakujących. Pozostałą część rozdziału zajmują obszerne testy w dwóch wariantach:

1. Tworzony jest syntetyczny model, w którym zmienna objaśniana jest kombinacją liniową 2-4 zmiennych ze zbioru z dodatkowym szumem gaussowskim.

2. Zmienną objaśnianą jest śmiertelność noworodków dla każdego kraju na przestrzeni lat.

Trudno byłoby mi przybliżyć tu wszystkie eksperymenty obliczeniowe zawarte w tym rozdziale, ponieważ jest ich bardzo wiele i wykonane są w sposób wyczerpujący. Doktorantka testuje oczywiście autorskie algorytmy z pracy (MI-ForwardStepwise, MIC-ForwardStepwise i MI-Lasso, tj. efektywną wersję algorytmu Lasso) w trzech modelach zawierających (odpowiednio) 2, 3 i 4 zmienne, przy stukrotnym powtórzeniu eksperymentu. Dla algorytmów regresji krokowej z użyciem indeksów wielowymiarowych badany jest odsetek przesianych zmiennych (jako funkcja zmiennych w modelu), jakość predykcji mierzona pierwiastkiem z błędu średniokwadratowego (również jako funkcja zmiennych w modelu) zarówno na zbiorze treningowym jak i testowym, wpływ kryterium stopu (poprzez zmierzenie częstotliwości wyboru modelu z daną liczbą zmiennych przy użyciu poszczególnych kryteriów stopu), a także porównanie wartości granicznych korelacji wynikających z kandydata x_0 z wartościami z kryterium stopu. Dla algorytmu MI-Lasso sprawdzono skuteczność przesiewania zmiennych, jakość predykcji na zbiorze treningowym i testowym, a także porównanie jakości przesiewania zmiennych z innymi metodami dla problemu Lasso przybliżonymi pod koniec rozdziału piątego, zarówno w wersjach nieiterowanych, jak i iterowanych. Autorka przygląda się również jakości rozwiązania kandydata x_0, przy okazji porównując dwie metody jego znajdywania w przypadku algorytmu MI-Lasso, co pozwala również porównać się z metodą OMP, która na takim kandydacie by poprzestała. Dalej, Doktorantka przygląda się algorytmowi SIS, porównując wartości korelacji dodawanych przez niego zmiennych (z wektorem zmiennej wyjściowej) z wartościami uzyskiwanych z wcześniej wymienionych metod. Przygląda się również odsetku znalezionych właściwych (lub prawie-właściwych) zmiennych i czasowi obliczeń, a także jakości predykcji wszystkich modeli. Dla modelu opartego na prawdziwej zmiennej objaśnianej przedstawione i analizowane są konkretne cechy dodawane w kolejnych krokach algorytmów.

Wyniki są w przeważającej mierze korzystne dla algorytmów Doktorantki i pokazują, że są one znacznie efektywniejsze pod względem czasu obliczeń od ich „klasycznych" odpowiedników z indeksów nie korzystających, a równocześnie co najwyżej minimalnie (lub wcale) tracą na jakości. Wyniki są miejscami imponujące – odsetek zwróconych (odsianych) zmiennych może stanowić mały ułamek wszystkich zmiennych (nawet rzędu 0.001%), a czasy obliczeń są o rzędy wielkości krótsze. Mnogość analiz robi duże wrażenie i pozwala zweryfikować wszystkie wybory (i ich wpływ) podjęte przy konstrukcji nowych algorytmów.

Ostatni z rozdziałów zawiera podsumowanie pracy.

3 Ocena wkładu oryginalnego

Rozprawa jest oparta na czterech artykułach, których współautorem jest Doktorantka:

- B. Żogała-Siudem, S. Jaroszewicz (2014): Fast stepwise regression on Linked Data. *Proc. of the 1st Workshop on Linked Data for Knowledge Discovery (LD4KD) co-located with ECML/PKDD '14*. Nancy, France

- B. Żogała-Siudem, S. Jaroszewicz (2015): Geometric approach to stepwise regression. *Computational Methods in Data Analysis. Information Technologies: Research and their Interdisciplinary Applications ITRIA 2015*, pp. 213-224.
- B. Żogała-Siudem, S. Jaroszewicz (2021): Fast stepwise regression based on multidimensional indexes. *Information Sciences*, 549:228-309
- B. Żogała-Siudem, S. Jaroszewicz (2021): Variable screening in Lasso homotopy algorithm with multidimensional indexing (w recenzji).

Wszystkie cztery prace są autorstwa wyłącznie Doktorantki wraz z promotorem, nie ma tu więc miejsca na wątpliwości dotyczących wkładu „zewnętrznych" autorów do prac. Ostatni z artykułów nie został opublikowany, ale stanowi podstawę dla jednego z ważnych podrozdziałów rozprawy, opisując modyfikację homotopijnego algorytmu budowy ścieżki regularyzacji dla algorytmu Lasso. Trzeci z artykułów został opublikowany w czasopiśmie o bardzo wysokim współczynniku *Impact Factor* wynoszącym 6,795 (najwyższy współczynnik 200 tzw. „punktów MNiSW"). Jest on jednocześnie rozszerzeniem pierwszego artykułu, który ukazał się uprzednio na konferencji. Wreszcie, drugi z artykułów to rozdział w książce wydanej przez IPI PAN. Dzięki obecności artykułu z *Information Science* uważam, że jest to dorobek w wystarczającym stopniu spełniający wymogi do finalizacji przewodu doktorskiego. Dodatkowo, po spojrzeniu na serwisy *Google Scholar* i *Orcid*, wyżej wymienione prace p. B. Żogały-Siudem nie stanowią całości dorobku naukowego, który jest obszerniejszy i nie został włączony do pracy w całości aby zachować pełną spójność tematyczną.

Rozprawa zawiera wiele oryginalnych wyników, które omówiłem już w poprzedniej części recenzji, w związku z tym tutaj wymienię kilka najważniejszych, moim zdaniem, osiągnięć:

- Konstrukcja algorytmów regresji krokowej przy użyciu indeksów wielowymiarowych, znacznie przewyższających oryginalne ich wersje pod względem szybkości obliczeń, bez straty na jakości. Algorytmy oparte są na wyczerpującej analizie teoretycznej z udowodnieniem poprawności użytych kryteriów odsiewania zmiennych, a w efekcie poprawności samego algorytmu.
- Konstrukcja efektywnego algorytmu homotopijnego dla problemu Lasso zwracającego całą ścieżkę regularyzacyjną przy użyciu indeksów wielowymiarowych, z taką samą dbałością o poprawność działania jak w punkcie poprzednim.
- Bardzo obszerne eksperymenty obliczeniowe z użyciem autorskich algorytmów na rzeczywistym zbiorze danych w modelach z syntetyczną jak i z rzeczywistą zmienną wyjściową, zawierające wyczerpujące porównanie algorytmów z istniejącymi rozwiązaniami względem czasu obliczeń, jakości predykcji i trafności dobranych zmiennych.
- Dodatkowe eksperymenty z jakością indeksów wielowymiarowych na zbiorze Eurostat oraz badanie zależności ułamka zwróconych punktów w funkcji progu na korelacji dla modelu losowego i danych z Eurostatu.

Krótko podsumowując, uznaję, że wymienione na wstępie pracy cele udało się Doktorantce w pełni osiągnąć.

4 Uwagi dyskusyjne

Nie kwestionując wartości całościowych wyników zawartych w rozprawie, chciałbym zgłosić poniżej kilka uwag, głównie w formie pytań bądź dyskusji.

- *Brak informacji o koszcie utworzenia indeksu.* Z założenia indeks odwrotny jest tworzony jednokrotnie, a później wielokrotnie używany na potrzeby analiz na obszernym zbiorze danych, ale interesujące byłoby określenie czasu obliczeniowego potrzebnego na utworzenie takiego indeksu i porównanie go z czasem „zaoszczędzonym" przez metody na nim oparte w późniejszych analizach. Pozwoliłoby to na oszacowanie zysku utworzenia takiego indeksu np. dla użytkownika, który zawczasu potrafi określić liczbę modeli, które zamierza wyznaczyć na zbiorze danych, a także dopełniłoby analizy czasu obliczeń wprowadzonych algorytmów.

- *Zależność działania algorytmów od liczby obserwacji* n – w części eksperymentalnej liczba ta jest trzymana jako stała, być może warto byłoby sprawdzić, czy podobne wnioski utrzymują się dla mniejszej lub większej liczby obserwacji (w szczególności: może to rzutować na dokładność i obszerność wyników zwracanych przez indeksy przybliżone).

- *Normalizacja zmiennych* – przez większość część pracy Autorka zakłada, że wszystkie zmienne są wstępnie znormalizowane. Nie niesie to prawdopodobnie dużego narzutu obliczeniowego, ponieważ musi zostać wykonane tylko raz, ale może potencjalnie zmieniać działanie algorytmu w stosunku do wersji bazujących na zmiennych oryginalnych (np. inaczej będzie się zachowywać funkcja kary w modelu Lasso; swoją drogą może to być działanie pożądane!). Czy istnieje prosta metoda pozwalająca na działanie na zmiennych znormalizowanych, a równocześnie odtworzenie działania algorytmu tak, jakby działał na zmiennych nieznormalizowanych?

  Przy okazji, nie zostało napisane, czy zmienna y w modelach syntetycznych (A, B, C) zostaje znormalizowana?

- *Kryteria stopu w oparciu o jakość predykcyjną* – z eksperymentów wynika, że algorytmy (zarówno wersje Autorki, jak i oryginalne bez użycia indeksów) dość szybko się „przeuczają", stąd być może jako kryterium stopu dałoby się użyć trafności predykcji na osobnym zbiorze walidacyjnym. Czy dałoby się tego typu miarę zaprząc do algorytmów regresji krokowej opartych na indeksach w podobny sposób, jak inne używane kryteria stopu?

- *Użycie kart graficznych do przyspieszenia obliczeń* – mój najbardziej spekulatywny komentarz dotyczący faktu, że wiele elementów w algorytmach opiera się tak naprawdę na wykonaniu operacji macierzowych dużej skali (np. wyznaczaniu korelacji dla dużego zbioru zmiennych na raz, co można przedstawić jako mnożenie macierzy i wektora, itp.), z którymi bardzo dobrze radzą sobie karty graficzne. Nasuwa się tu pytanie, czy pozwoliłoby to na zwiększenie wydajności oryginalnych wersji algorytmów (nie używających indeksów) lub może również nowych algorytmów wprowadzonych w pracy?

5 Konkluzja końcowa

Rozprawę oceniam bardzo dobrze, co w powyżej recenzji podkreśliłem wielokrotnie. Problemy badawcze, z którymi zmierzyła się Doktorantka, są ambitne i istotne dla postępu w dziedzinie selekcji zmiennych i rozwiązywania problemów o ekstremalnie dużej liczbie predyktorów. Sama rozprawa charakteryzuje się wysokim poziomem merytorycznym i zawiera interesujące metody i rezultaty, zarówno o charakterze teoretycznym, jak i praktycznym. W mojej opinii Doktorantka wykazał się znakomitymi umiejętnościami prowadzenia badań naukowych. Wszystkie postawione w rozprawie cele zostały osiągnięte.

W związku z tym **rozprawę oceniam jako spełniającą wymogi stawiane pracom doktorskim i wnoszę o dopuszczenie mgr Barbary Żogały-Siudem do dalszych etapów przewodu doktorskiego, równocześnie sugerując wyróżnienie pracy.**

dr hab. inż. Wojciech Kotłowski

Wojciech Kotłowski