dr hab. inż. Małgorzata Bogdan
Instytut Matematyki
Uniwersytet Wrocławski

Wrocław 11.10.2016

RECENZJA ROZPRAWY DOKTORSKIEJ

Tytuł rozprawy: Delete or Merge Regressors algorithm

Autor rozprawy: mgr Agnieszka Prochenka

Promotor rozpawy: dr hab. Piotr Pokarowski

Cel i charakter rozprawy

Rozprawa dotyczy identyfikacji optymalnego modelu (uogólnionej) regresji liniowej w sytuacji gdy część predykatorów ma charakter jakościowy. W tej sytuacji redukcja rozmiaru modelu polega na identyfikacji zerowych współczynników regresji a także na pogrupowaniu kategorii zmiennych jakościowych w klasy, wewnątrz których nie ma zróżniowania wartości zmiennej zależnej. W rozprawie zaproponowano konkretne algorytmy identyfikacji takich optymalnych modeli i udowodniono ich zgodność w sytuacji gdy liczba kolumn p w pełnej macierzy eksperymentu jest mniejsza niż liczba wierszy n. Zaproponowano także rozszerzenia tych algorytmów na przypadek $p >> n$, które zweryfikowano za pomocą obszernych symulacji komputerowych.

Struktura rozprawy

Rozprawa doktorska składa się z 6 rozdziałów. W pierwszym rozdziale krótko przedstawiono cel, zakres i tezy pracy. Rozdział drugi zawiera krótkie wyjaśnienie problemu identyfikacji grup "podobnych" poziomów zmiennych jakościowych i wprowadzenie matematycznej notacji. W rozdziale tym można znaleźć także krótkie informacje na temat istniejących narzędzi do analizy takich danych. Rozdział trzeci dotyczy algorytmu DMR4lm dla modelu regresji liniowej. Zawiera opis algorytmu a także dowód jego zgodności. Rozdział czwarty ma zbliżoną strukturę i dotyczy algorytmu modelu DM4glm dla uogólnionej regresji liniowej. Rozdział piąty i szósty zawierają wyniki obszernych analiz rzeczywistych zbiorów danych i symulacji komputerowych.

Ocena pracy doktorskiej

Rozprawa doktorska mgr. Agnieszki Prochenki dotyczy bardzo ważnego zagadnienia redukcji wymiaru przy dopasowaniu modelu liniowego. Taka redukcja wymiaru ma wartości poznawcze (identyfikacja istotnych czynników wpływających na zadaną cechę) a także umożliwia redukcję wariancji estymatorów istotnych parametrów i korzystnie wpływa na własności predykcyjne dopasowanych modeli. W pracy podjęto trudny temat redukcji wymiaru poprzez identyfikację "podobnych" (z punktu widzenia wartości zmiennej zależnej) poziomów zmiennych czynnikowych. Problem jest trudny ze względu na znaczną liczbę możliwych partycji poziomów czynnika, które należałoby przejrzeć w celu identyfikacji optymalnego modelu. Zaproponowany algorytm rozwiązuje ten problem poprzez zastosowanie algorytmu klastrowania hierarchicznego, w wyniku którego dostajemy ciąg interesujących zagnieżdżonych modeli. Optymalny model z tego ciągu jest następnie wybierany poprzez zastosowanie Uogólnionego Kryterium Informacyjnego. Algorytm tworzący ciąg modeli zagnieżdżonych wymaga wyznaczenia nieobciążonych estymatorów parametrów modelu liniowego i działa w sytuacji gdy $p < n$. W przeciwnym wypadku liczba zmiennych jest wstępnie zredukowana poprzez zastosowanie grupowego LASSO. W pracy sformułowano warunki przy których DMR4lm and DMR4glm są zgodne i udowodniono odpowiednie twierdzenia. Ponadto działanie obu algorytmów zilustrowano obszernymi symulacjacmi i analizą danych rzeczywistych. Wyniki dotyczące DMR4lm zostały opublikowane w artykule w renomowanym czasopiśmie *Electronic Journal of Statistics*. W mojej ocenie wyniki te są bardzo wartościowe. Klasyczne wyniki

z tej dziedziny dotyczą zgodności kryteriów informacyjnych, bez wskazania w jaki sposób można takie kryterium zoptymalizować. Udowodnienie wyników dotyczących zgodności konkretnego algorytmu porządkującego zmienne i dokonującego ich wyboru jest znacznie bardziej złożone.

Jak już wspomniałam, problem rozważany przez autorkę jest złożony technicznie i jego dobre wyjaśnienie wymaga sporej pracy edytorskiej. W moim odczuciu tego edytorskiego wysiłku brakuje w rozprawie doktorskiej. Wydaje mi się, że rozprawa by znacznie skorzystała gdyby połowę miejsca wykorzystanego na symulacje przeznaczyć na bardziej uporządkowany i kompletny opis stosowanych rozwiązań i dogłębną dyskusję wyników teoretycznych. Listę uwag krytycznych dotyczących rozprawy umieściłam na kolejnych stronach. Zauważone usterki nie wpływają na moją ocenę poprawności zastosowanych przez autorkę metod i pozytywną ocenę rozprawy.

Moim zdaniem rozprawa doktorska mgr Agnieszki Prochenki spełnia wymogi ustawy o stopniach i tytułach naukowych i wnioskuję o dopuszczenie jej do publicznej obrony.

Z wyrazami szacunku,
Małgorzata Bogdan M. Bog

Uwagi krytyczne/tematy do dyskusji

1. Konstrukcja ciągu modeli zagnieżdżonych w dużej mierze jest zdeterminowana działaniem algorytmu klastrowania hierarchicznego. Moim zdaniem praca by zyskła gdyby temu tematowi przeznaczono osobny podrozdział. Algorytm hierarchiczny pojawia się w opisie DMR4lm wraz z grupą pojęć (jak np. "cutting heights"), które są wyjaśnione dopiero w kolejnych rozdziałach. Bardzo utrudnia to czytanie pracy.

2. Założenia twierdzeń o zgodności zależą od wartości parametrów δ_F i δ_T, które powinny dążyć do nieskończoności. Zdecydowanie przydałaby się dyskusja tego warunku - np. przetłumaczenie na język wielkości współczynników regresji i własności macierzy eksperymentu.

3. Zgodność udowodniona jest przy bardzo ograniczających założeniach na współczynnik kary r_n w GIC (Generalized Information Criterion). W szczególności zakłada się, że $p_n = o(r_n)$, co wyklucza przypadek $r = c \log p$. Dlatego nie za bardzo rozumiem dlaczego w symulacjach i analizie danych rzeczywistych użyto $r = c \log p$.

4. Warunek na GIC podany we Wniosku 1 jest bardzo restrykcyjny. Czy nie można by go osłabić przy dodatkowym założeniu ograniczającym np. maksymalną liczbę poziomów zmiennych jakościowych ?

5. Opis algorytmu DMRnet4lm jest dla mnie nieczytelny. Z przyjemnością wysłucham tego opisu na obronie.

6. W symulacjach autorzy stwierdzają,że ich metoda daje błąd predykcji porównywalny z błedami metod bezpośrednio wykorzystujących LASSO itp. ale przy mniejszej liczbie predykatorów. Jest to fakt dość oczywisty, bo metody które ściągają estymatory do zera generalnie wymagają niższych progów aby uzyskać dobrą predykcję. Moim zdaniem warto byłoby porównać DMR4lm z innymi metodami wykorzystującymi estymatory "nieobciążone" jak np. z metodą regresji wstępującej. W przypadku bardzo rzadkich modeli i dużego p takie podejście mogłoby znacznie zredukować wariancję estymatorów i poprawić własności uzyskanego ciągu modeli zagnieżdżonych. Poza tym warto byłoby porwnać jak zadziała GIC zastosowane do ciągu modeli zagnieżdżonych uwtorzonych w oparciu o MCP lub CAS-ANOVA.

7. Brakuje mi sugestii dotyczących konkretnego wyboru GIC (tzn. wyboru r). Jeżeli r ustalane jest w oparciu o walidację krzyżową to to już nie jest GIC (tylko walidacja krzyżowa).

8. W Twierdzeniu 2 zakładamy, że t i p są skończone, więc pierwszy warunek w założeniu nie zachodzi.

Drobne uwagi redakcyjne

Paragraf 2.2, trzecia linia - liczba ograniczeń to $p - q$ a nie q.

Strona 11 - dlaczego zakładamy, że $t \leqslant p - 2$? Czy metoda nie jest w stanie zidentyfikować pełnego modelu ?

Strona 23 - Jak duże są typowo δ_T i δ_F (tzn. jak duże może być r)?