dr hab. Wojciech Rejchel, prof. UMK
Wydział Matematyki i Informatyki
Uniwersytet Mikołaja Kopernika

Toruń, 29.12.2022

Recenzja rozprawy doktorskiej mgr Małgorzaty Łazęckiej

„Properties of information-theoretic measures of conditional dependence"

Testy warunkowej niezależności odgrywają ważną rolę w wielu problemach statystyki i uczenia maszynowego, na przykład w selekcji cech, rozpoznawaniu struktury modeli graficznych czy poszukiwaniu przyczynowości. Dlatego tematyka ta jest aktualnie popularna i intensywnie badana. Problematyka rozważana w ocenianej rozprawie doktorskiej również wpisuje się w ten nurt.

Mgr Małgorzata Łazęcka słusznie zauważa, że warunkowa informacja wzajemna (z ang. conditional mutual information, CMI), czyli jedno z głównych pojęć teorii informacji, ma wiele własności, dzięki którym może być używana do badania warunkowej niezależności między zmiennymi. W rozprawie ograniczono się do zmiennych dyskretnych i wykazano, że znalezienie asymptotycznego rozkładu estymatora CMI nie jest zbyt kłopotliwe. Niestety metoda ta napotyka na znaczące trudności estymacyjne, co wymusza podejście oparte na uproszczonych estymatorach CMI. Teoretyczne odpowiedniki tych ostatnich nie posiadają użytecznych w testowaniu warunkowej niezależności własności CMI, co sprawia, że ich analiza statystyczna jest znacznie trudniejsza. W rozprawie problemy te zostały rozwiązane, używając metod wielokrotnego próbkowania. Ważną część rozważań stanowi analiza numeryczna, w której zbadano własności zaproponowanych metod. Zobrazowane zostały zarówno ich zalety, jak i wady.

Omówienie rozprawy doktorskiej

Rozdział pierwszy rozpoczynają definicje podstawowych pojęć teorii informacji (entropia, informacja wzajemna i ich warunkowe wersje) oraz niektóre ich własności. W części 1.1.2 omówiono użycie tych narzędzi w selekcji cech i uzasadniono fakt, że kluczowym obiektem do badania jest warunkowa informacja wzajemna. Następnie zdefiniowano informację interakcyjną i jej związki z entropią i informacją wzajemną. Podsumowaniem tej części rozprawy jest Twierdzenie 1.2.10, czyli rozwinięcie Möbiusa CMI, które pozwala przedstawić ją jako sumę informacji interakcyjnych kolejnych stopni. Zatem „kompletna estymacja" CMI wymaga przybliżenia wszystkich informacji interakcyjnych (do

pewnego rzędu), co w praktyce zwykle skazane jest na porażkę ze względu na niewystarczającą liczbę obserwacji. Sprawia to, że stosowane kryteria ograniczają się do rozwinięć CMI rzędu dwa bądź trzy.

W części 1.4 zbadano asymptotyczne rozkłady estymatorów i ich użyteczność w testowaniu warunkowej niezależności. W Lemacie 1.4.2 przytoczono wcześniej znaną własność estymatora CMI, która mówi, że przy warunkowej niezależności właściwie przeskalowany estymator CMI ma asymptotycznie rozkład χ^2 z odpowiednią liczbą stopni swobody. Sytuacja komplikuje się w przypadku estymatorów uciętych (Twierdzenie 1.4.6), gdyż tutaj postać rozkładu granicznego zależy od nieznanej w praktyce wielkości $D^T \Sigma D$, gdzie D jest gradientem funkcji kryterialnej, a Σ to asymptotyczna wariancja estymatora uzyskanego metodą podstawiania częstości (albo krócej, z ang., *plug-in*). Jeśli wielkość ta jest dodatnia, to asymptotycznie mamy rozkład normalny z klasycznym rzędem $\sqrt{n}$. W przeciwnym przypadku otrzymujemy modyfikację rozkładu χ^2 z rzędem n. Już w tym momencie można zauważyć, że ucinanie estymatorów CMI pociąga za sobą konieczność ich centrowania, co będzie generowało pewne problemy w praktyce. Twierdzenie 1.4.6 elegancko uogólnia fakty znane z prac [18, 20].

Estymatory oparte na rozwinięciu rzędu dwa są jeszcze dokładniej omówione w Lemacie 1.4.10 oraz Twierdzeniu 1.4.11 przy dodatkowym założeniu, że Y jest binarne. Uzasaniono tu użycie tych estymatorów do testowania serii *pojedynczych* hipotez o warunkowej niezależności.

Wspomniane powyżej kłopoty z estymatorami CMI oraz ich uciętymi wersjami zostały rozwiązane w rozdziale drugim, używając metod wielokrotnego próbkowania (z ang. bootstrap). Rozważono cztery metody bootstrapowe: *pełne* próbkowanie z rozkładu odpowiadającemu warunkowej niezależności, jego *częściowe* odpowiedniki podobne do metody *knockoffs* [8] oraz wersję permutacyjną. W każdej sytuacji wyznaczono rozkład asymptotyczny estymatora typu *plug-in*: Lematy 2.1.1 - 2.1.4. Następnie w części 2.2.1 uzasadniono, że asymptotyczny rozkład estymatora CMI (otrzymany w Lemacie 1.4.2) jest taki sam, jak jego czterech bootstrapowych wersji. Pozwala to na konstrukcję asymptotycznych testów warunkowej niezależności, które wyprowadzone są w części 2.2.2. Ponadto dla dwóch metod wielokrotnego próbkowania, które *zachowują wymienialność*, otrzymano testy nieasymptotyczne.

Na koniec rozdziału drugiego podjęto próbę zbadania asymptotycznego rozkładu bootstrapowej wersji estymatora uciętego. W Lemacie 2.3.1 udało się uzyskać odpowiednik Twierdzenia 1.4.6, gdy rozkład asymptotyczny jest normalny.

Rozdział trzeci zawiera wyniki eksperymentów numerycznych, w których badano własności zaproponowanych metod na danych symulowanych. Rozważono zarówno estymatory CMI, jak i ich wersje ucięte. Te pierwsze były reprezentowane przez estymator oparty na rozkładzie granicznym z Lematu 1.4.2, a także jego modyfikacje bazujące na wielokrotnym próbkowaniu (jak w pracy [41] lub metody porównujące estymator CMI z kwantylami jego wersji bootstrapowych). Rozważane estymatory ucięte rzędu dwa bądź trzy to, niedostępny w praktyce,

estymator z Twierdzenia 1.4.6, jego uproszczona wersja bazująca na rozkładzie χ^2 oraz estymatory związane z wielokrotnym repróbkowaniem. Analizowano głównie prawdopodobieństwo błędu pierwszego i drugiego rodzaju procedur.

Wnioski z przeprowadzonych badań zawarto w części 3.3 zatytułowanej „Podsumowanie eksperymentów", ale wydaje mi się, że ten podrozdział można by zatytułować „Podsumowanie rozprawy". Znajdziemy tam wyczerpujące omówienie problemów związanych z testowaniem warunkowej niezależności (oraz zaproponowanych rozwiązań): począwszy od kłopotów z estymacją CMI, a następnie z estymatorami uciętymi (miejsce ucięcia, *dwoistość* rozkładu granicznego, konieczność centrowania itd.). Nie zabrakło również rozważań dotyczących użycia wyników asymptotycznych do *skończonych* zbiorów danych.

Uwagi

1) strona 16: informacja wzajemna oraz warunkowa informacja wzajemna są mylnie nazwane miarami siły zależności i warunkowej zależności. O obiektach tych należy raczej myśleć jak o *indykatorach* niezależności, które jednakże nie opisują siły zależności (analogicznie, to nie kowariancja lecz jej *unormowana* wersja opisuje siłę zależności). Dla przykładu, $I(X,Y) = H(X)$ dla $X = Y$, czyli informacja wzajemna między zmienną X a nią samą może być mała, jeśli X ma małą entropię. Dlatego bardziej właściwe byłoby używanie unormowanej informacji wzajemnej. Jednakże fakt ten nie wpływa na poprawność rozumowań zawartych w rozprawie, co można zauważyć, na przykład, przyglądając się testom opisanym w częściach 2.2.2 oraz 2.2.3,

2) kryteria selekcji cech oparte na informacji wzajemnej wymagają, z oczywistych względów, ograniczenia liczności cech rozważanych kandydatów (w części 1.1.2 jest to warunek $|T| = k$). Wydaje się, że ten niewygodny warunek mógłby być pominięty, używając kryteriów opartych na informacji wzajemnej z karą (analogicznie jak w kryterium Akaike bądź Schwarza). Czy tego typu (bądź podobne) procedury były już badane? Jeśli tak, to z jakim skutkiem,

3) standardowe metody selekcji cech zwykle zakładają pewien związek między zmienną zależną a predyktorami, na przykład dla zmiennej binarnej często jest to zależność logistyczna. W oczywisty sposób zawęża to ogólność rozważań. Jednakże zaletą tego podejścia jest możliwość sprawdzenia jakości otrzymanego estymatora przy pomocy, na przykład, predykcji na nowych obiektach (*dobra* selekcja powinna pociągać *dobrą* predykcję). Co mogłoby odgrywać rolę tego typu „sprzężenia zwrotnego" w kryteriach opartych na informacji wzajemnej?

4) w dowodzie Lematu 2.1.1 rozumowanie prowadzone jest warunkowo (przy ustalonych $(X_i, Y_i, Z_i)_{i\geq 1}$), więc $\hat{p}_{ci}$ jest ustalonym ciągiem *liczbowym*. Zatem niezbyt właściwe jest stwierdzenie, że $\hat{p}_{ci}$ jest zbieżne *prawie wszędzie*. Ono jest zbieżne albo nie jest. W dowodzie można przyjąć, że jest zbieżne, bo w założeniu lematu napisane jest *dla prawie wszystkich ciągów* $(X_i, Y_i, Z_i)_{i\geq 1}$, czyli już na początku można ograniczyć się do tych ciągów danych, które dają zbieżność $\hat{p}_{ci}$. Podobnie dla $\hat{\Sigma}_{-K}$, dowodu Lematu 2.1.2 itd.,

5) szkoda, że w części eksperymentalnej rozprawy nie zbadano skuteczności rozważanych metod w rozpoznawaniu struktury grafów bądź w selekcji cech, a także nie porównano ich z innymi procedurami (na przykład używającymi bardziej restrykcyjnych założeń). Zwłaszcza przypadek (dosyć) dużych zbiorów danych byłby ciekawy.

Konkluzja

Uważam, że rozprawa zawiera nowe i interesujące wyniki dotyczące testowania warunkowej niezależności. Dowody twierdzeń wymagały od Kandydatki opanowania dość zaawansowanych narzędzi matematycznych i statystycznych, a także sumienności, dokładności i rachunkowej wprawy. Ponadto umiejętne zaproponowanie nowych metod niewątpliwie bazowało na dobrej znajomości trudnych działów (bądź ich fragmentów), jakimi są teoria informacji czy wielokrotne próbkowanie.

Nie mam wątpliwości, że przedstawiona rozprawa doktorska spełnia ustawowe i zwyczajowe wymagania stawiane rozprawom doktorskim w dyscyplinie matematyka. Wnoszę o dopuszczenie mgr Małgorzaty Łazęckiej do dalszych etapów przewodu doktorskiego.

Wojciech Rejchel