prof. dr hab. inż. Małgorzata Bogdan
Instytut Matematyki
Uniwersytet Wrocławski

Wrocław 22.11.2021

RECENZJA ROZPRAWY DOKTORSKIEJ

Tytuł rozprawy: Maximal a Posteriori Partition in Nonparametric Bayesian Mixture Models with applications to Clustering Problems

Autor rozprawy : mgr Łukasz Rajkowski

Promotorzy rozprawy: prof. dr hab. Wojciech Niemiro i dr John Noble

Cel i charakter rozprawy

Rozprawa doktorska mgr. Łukasza Rajkowskiego dotyczy analizy skupień, która jest jednym z klasycznych i podstawowych zagadnień statystycznego uczenia bez nadzoru. Głównym celem analizy skupień jest podział zbioru danych na naturalnie różne podzbiory. Jedna z podstawowych technik polega na dopasowaniu do danych modelu mieszanek gaussowskich (Gaussian Mixture Model) z ustaloną liczbą składowych. W kolejnym kroku zwykle dokonuje się wyboru optymalnej liczby składowych za pomocą różnych kryteriów, jak np. Bayesowskiego Kryterium Informacyjnego. W swoim doktoracie mgr Rajkowski dokonuje szczegółowej analizy nieparametrycznej bayesowskiej wersji tej metody, w której nie ogranicza się maksymalnego rozmiaru partycji a różne różnoliczne partycje bezpośrednio porównuje się za pomocą ich prawdopobieństw aposteriori. W takim bayesowskim modelu nieparametrycznym rozkład apriori na nieskończone ciągi p-stw przynależności do poszczególnych skupień opisuje się za pomocą procesu Dirichleta a rozkład p-stwa na nieskończonym zbiorze możliwych partycji zadany jest tzw. Procesem Chińskiej Restauracji. Punktem wyjściowym doktoratu mgr. Rajkowskiego jest artykuł [1], w którym zauważono, że liczba skupień wyestymowana w oparciu o ten model nie jest zgodnym estymatorem rzeczywistej liczby składowych

mieszanki generującej dane. Dogłębna analiza matematyczna zaproponowana przez mgr. Rajkowskiego stara się wyjaśnić przyczyny tego zjawiska oraz proponuje rozwiązanie, które być może wskazuje drogę do konstrukcji zgodnego estymatora liczby skupień. Częściowe wyniki rozprawy zostały opublikowane w samodzielnym artykule w renomowanym czasopiśmie *Bayesian Analysis* (140pt MNiSW).

Struktura i treść rozprawy

Rozprawa doktorska napisana jest w języku angielskim. Składa się z 5 rozdziałów i dodatku.

W pierwszym rozdziale wprowadzono ogólny model mieszanek bayesowskich oraz jego nieparametryczne wersje wykorzystujące procesy Dirichleta i Pitmana-Yora. Rozdział ten zawiera także opis sprzężonych rodzin wykładniczych i szczególnych modeli stosowanych w rozprawie. Są to model normalno-normalny (NN), w którym macierz kowariancji wewnątrz skupień jest znana, model NIW (Normal-Inverse-Wishart), w którym rozkład apriori na macierze kowariancji wewnątrz skupień jest odwrotnym rozkładem Wisharta i nowy model NIG (Normal-Inverse-Gamma), który poprzednio w literaturze występował jedynie w przypadku jedno-wymiarowym.

W Rozdziale 2 przedstawiono szereg interesujących wyników dotyczących własności partycji maksymalizujących p-stwo aposteriori (Maximum A Posteriori, MAP) w powyższych modelach. W szczególności udowodniono, że partycja MAP w sprzężonym wykładniczym modelu mieszanek bayesowskich jest separowalna za pomocą liniowych funkcjonałów statystyk dostatecznych. Ten wynik jest istotnym rozszerzeniem wyniku z pracy [2], w którym udowodniono liniową separowalność różnych elementów partycji w modelu NN.

W Rozdziale 2 szczegółowo przeanalizowano zachowanie rozkładów aposteriori na partycjach indukowanych przez dowolny rozkład P generujący dane. W szczególności, w ogólnym sprzężonym modelu mieszanek bayesowskich wyznaczono asymptotyczną granicę ciągu nieunormowanych prawdopodobieństw aposteriori dla dowolnej P indukowanej partycji gdy obserwacje są niezależnymi wektorami losowymi z rozkładu P. Istotną składową tego przybliżenia jest funkcja Δ_P, której maksymalizacja prowadzi do maksymalizacji asymptotycznego p-stwa aposteriori. Wyznaczono dokładną postać tej funkcji dla trzech normalnych modeli mieszanek Bayesowskich wprowadzonych w Rozdziale 1 a następnie wykorzystano ją do konstrukcji asymptotycznie optymalnej partycji odcinka [0,1] indukowanej przez rozkład jednostajny na tym

odcinku. W wyniku tej analizy ustalono, że w modelu NN optymalna partycja indukowana składa sę z odcinków o równej długości, a liczba tych odcinków zależy od założonej wariancji wewątrz skupień. Z kolei w przypadku modelu NIW stwierdzono, że dowolna partycja odcinka [0,1] daje tą samą wartości funkcji Δ_P, co, zdaniem autora rozprawy, jest zjawiskiem niepożądanym. W moim odczuciu jest to jednak wynik oczekiwany, gdyż rozkład jednostajny można przedstawić jako mieszankę rozkładów jednostajnych na elementach dowolnej partycji odcinka [0,1].

W Rozdziale 3 szczegółowo przedstawiono wyniki z pracy [2] dotyczące asymptotyki partycji MAP w modelu NN. W szczególności pokazano, że jeżeli ciąg obserwacji spełnia warunek ograniczoności próbkowej wariancji to rozmiar najmniejszego skupienia rośnie proporcjonalnie do liczby obserwacji. Ponadto, jeżeli ciąg obserwacji jest ograniczony to liczba skupień przecinających dowolną kulę pozostaje ograniczona gdy liczba obserwacji dąży do nieskończoności. Wyniki te generalizują się do sytuacji gdy ciąg obserwacji jest generowany losowo z rozkładu o skończonym czwartym momencie lub ograniczonym nośniku, odpowiednio.

W mojej ocenie najtrudniejsze i najciekawsze wyniki rozprawy znajdują się w części 3.3 w której udowodniono, że w przypadku gdy dane są generowane z ciągłego rozkładu P o ograniczonym nośniku, ciągi wypukłych otoczek rożnych skupień partycji MAP w modelu NN zbiegają do skończonych P-indukowanych partycji, które maksymalizują Δ_P^{NN}. Dowód tego twierdzenia jest złożony i wymaga wiedzy z topologii i analizy zbiorów i funkcji wypukłych. Wynik jest też ważny ze statystycznego punktu widzenia, bo oznacza, że asymptotyczne własności partycji MAP można badać analizując funkcję Δ_P^{NN}. W tym rozdziale znajduje się też twierdzenie, które pokazuje, że dla ustalonego rozkładu P liczba skupień zależy od założonej macierzy kowariancji wewnątrz skupień. Następnie mgr Rajkowski stwierdza, że ten fakt można powiązać z brakiem zgodności estymatora liczby skupień. W mojej ocenie ta krytyka nie jest zasadna, ponieważ model NN zakłada, że Σ_0 jest znaną prawdziwą macierzą kowariancji i zgodność należy badać przy tym założeniu. Będę wdzięczna za komentarz na obronie na temat ewentualnego zastosowania Twierdzenia 3.21 do analizy zgodności w przypadku prawidłowej specyfikacji modelu.

W rozdziale 4 znajduje się analiza partycji MAP dla modelu NIW. Autor stwierdza (bez szerszego komentarza, będę wdzięczna za dyskusję przy obronie), że klasyczny model NIW ma zbyt dużo parametrów i proponuje model z regularyzacją. W modelu tym rozkład apriori na macierz kowarian-

cji wewnątrz skupień jest ściągany w kierunku pewnej ustalonej macierzy Σ_0, a współczynnik ściągający jest proporcjonalny do rozmiaru próby, ze współczynnikiem proporcjonalości λ. Tak więc w bardzo istotny sposób zostaje ograniczona elastyczność w estymacji macierzy kowariancji. Autor wyznacza odpowiednią funkcję $\Delta_{P,\lambda}$ i pokazuje, że gdy λ dąży do nieskończoności funkcja ta zbiega (z dokładnością do stałej) do funkcji Δ_P^{NN}. Tak więc zaproponowany model pozwala na ciągłe przejście między modelem NN a modelem NIW. Ponadto autor pokazuje, że dla dowolnego $\lambda > 0$ model ten zachowuje własności modelu NN w sensie ograniczonej granicznej liczby skupień i liniowego wzrostu liczby obserwacji w najmniejszym skupieniu. Autor proponuje zastosowanie empirycznego odpowiednika funkcji $\Delta_{P,\lambda}$ jako kryterium do porównywania jakości różnych partycji.

W rozdziale 4 pewne wątpliwości budzi staranie autora o uzyskanie ograniczonej liczby skupień gdy $n \to \infty$. Generalnie idea metod nieparametrycznych polega na tym, że mają one możliwość dobrego przybliżenia bardzo złożonych rozkładów (np.o nieskończonej liczbie skupień) jeżeli pozwala na to rozmiar próby. Tak więc wydaje się właściwe, żeby możliwa do uzyskania liczba skupień rosła w funkcji liczby obserwacji (np. jak pierwiastek z liczby obserwacji).

W rozdziale 5 znajdują się wyniki symulacji i analiza danych rzeczywistych ilustrujące działanie zaproponowanej przez autora funkcji kryterialnej. Rozdział ten wydaje się być słabszym ogniwem rozprawy. Autor podaje wynik tylko jednej symulacji i stosuje algorytm K-średnich do wyznaczenie "najlepszej partycji" o ustalonym wymiarze. Jak dobrze wiadomo wynik algorytmu K-średnich zależy od punktu startowego i w praktyce wykonuje się go wielokrotnie startując z różnych, zwykle losowych, punktów. Trochę zaskakuje fakt, że nie dokonano porównań ze standardowymi technikami analizy skupień opartymi na skończonych mieszankach rozkładów gaussowskich. Dodatkowo, trudność zagadnienia zależy od naturalnej rozdzielności skupień, tzn. od stosunku odległości między średnimi dla poszczególnych skupień a wariancją wewnątrz skupień. Przydałaby się analiza porównawcza dla kilku przykładów o różnym stosunku sygnału do szumu. Jeszcze innym zagadnieniem jest wybór parametru Σ_0. Wydaje się, że zastąpienie macierzy kowariancji wewnątrz skupień przez próbkową macierz kowariancji w pełnym zbiorze danych prowadzi do dużego przeszacowania wariancji wewnątrz skupień, co w efekcie prowadzi do zbyt małej liczby skupień dla większych wartości parametru λ.

Ocena pracy doktorskiej

W moim odczuciu jest to niewątpliwie wyróżniająca się rozprawa doktorska. Poziom matematyczny rozprawy znacznie przewyższa większość prac doktorskich ze statystyki. Zastosowane techniki dowodowe są pomysłowe i wymagały biegłości w różnych działach matematyki. W stosunku do klasycznych statystycznych wyników, które zwykle dowodzi się przy konkretnych założeniach na rozkład generujący dane, wyniki mgr Rajkowskiego dotyczące modelu NN zawierają kompletny opis asymptotycznego zachowania partycji MAP dla dowolnego rozkładu, co umożliwia uzyskanie wyników dotyczących zgodności w przypadku błędnej specyfikacji modelu. O wadze tych wyników świadczy publikacja w renomowanym czasopiśmie. Wyniki dotyczące modelu NIW w mojej ocenie wymagają jeszcze dopracowania. W szczególności nie jest dla mnie jasne czy zaproponowana regularyzacja jest właściwa w kontekście identyfikacji skupień w bardzo dużych zbiorach danych. Nie zmienia to jednak mojej bardzo wysokiej oceny rozprawy. W konkluzji stwierdzam, że:

Rozprawa doktorska mgr. Łukasza Rajkowskiego spełnia wymogi ustawy o stopniach i tytułach naukowych i wnioskuję o jej dopuszczenie do publicznej obrony. Proponuję również wyróżnienie rozprawy.

Z wyrazami szacunku,

Małgorzata Bogdan

Bibliografia

[1] J. W. Miller and M. T. Harrison, "Inconsistency of Pitman-Yor process mixtures for the number of components", Journal of Machine Learning Research, 15(1):3333– 3370, 2014.

[2] Ł. Rajkowski, "Analysis of the maximal a posteriori partition in the Gaussian Dirichlet Process Mixture Model", Bayesian Analysis, 14(2):477–494, 2019.