mgr inż. Małgorzata Żak-Szatkowska
Politechnika Wrocławska, Instytut Matematyki i Informatyki

Kryterium wyboru modelu dla regresji logistycznej

Wśród zagadnień predykcyjnych spotykanych w analizie danych często pojawia się problem klasyfikacji, czyli przypisania badanych obiektów do ustalonych klas. Decyzję o przypisaniu obiektu do jednej z klas podejmujemy na podstawie zestawu cech tych obiektów, tzw. predyktorów. Klasy, w zależności od ich liczby i rodzaju, możemy opisać zmienną binarną, z rozkładu Poissona lub inną zmienną dyskretną.

Jedną z metod stosowanych do predykcji zmiennych binarnych jest regresja logistyczna. Budując model regresji decydujemy, ile i które zmienne do niego włączymy. Wiadomo, że dopasowanie modelu na ogół poprawia się wraz z kolejnym dodanym regresorem, jednak zbyt rozbudowane modele mogą być przeuczone, tzn. zbyt dobrze dopasowane do zbioru danych, na którym powstały, a tym samym nieprzydatne w dalszych badaniach. Aby ograniczyć liczbę regresorów, możemy korzystać z kryteriów wyboru modelu, np. AIC lub BIC, które porównując modele między sobą biorą pod uwagę również ich wymiar.

W naszych badaniach zajmujemy się zagadnieniem lokalizacji genów. Chcemy określić, które geny wpływają na pewną cechę, np. powodują występowanie określonej choroby lub schorzenia. Rolę predyktorów pełnią tutaj genetyczne markery oraz interakcje między nimi, a ich łączna liczba może sięgać nawet kilku tysięcy. W sytuacji, gdy szukamy zaledwie kilku silnych genów wśród tak wielu potencjalnych zmiennych objaśniających, standardowe kryterium BIC przeszacowuje liczbę zmiennych w modelu. Dlatego w badaniach korzystamy również ze zmodyfikowanej wersji kryterium, mBIC [1], która jest przystosowana do opisanej sytuacji.

Oba kryteria, BIC oraz mBIC, opierają się na statystyce testu ilorazu wiarygodności (LRT). W przypadku regresji logistycznej stosowanie LRT jest często czasochłonne, ponieważ znalezienie maksimum funkcji wiarygodności wymaga stosowania metod numerycznych. Aby ułatwić i usprawnić obliczenia, wykorzystaliśmy w kryterium statystykę testu Rao [2], która, podobnie jak LRT, ma asymptotyczny rozkład χ^2. Wyniki uzyskane w obu przypadkach (kryterium ze statystyką LRT oraz Rao) porównaliśmy z wynikami dla zwykłej regresji liniowej.

Porównanie kryteriów BIC oraz mBIC potwierdza wcześniejsze wyniki: modele wybierane przez klasyczne BIC są zbyt rozbudowane. Rezultaty przy zastosowaniu LRT oraz statystyki testu Rao są porównywalne. Za stosowaniem testu Rao przemawia jego prostota oraz szybkość obliczeń. Metodę tę z powodzeniem zastosowaliśmy również dla danych z rozkładu Poissona i regresji log-liniowej.

Literatura

[1] M. Bogdan, J. K. Ghosh, R. W. Doerge, *Modifying the Schwarz Bayesian Information Criterion to locate multiple interacting quantitative trait loci*, Genetics 167 (2004), 989–999.

[2] W. C. M. Kallenberg, T. Ledwina, *Data driven smooth tests when the hypothesis is composite*, J. Amer. Statist. Assoc. 92 (1997), 1094–1104.