Streszczenie

Praca doktorska jest poświęcona problemowi wyboru modelu dla regresji liniowej i logistycznej kiedy predyktorami są zarówno zmienne ilościowe, jak i jakościowe, czyli czynniki. W literaturze opisane są dwie grupy algorytmów, które podejmują ten problem. Do pierwszej z nich należy dobrze znane grupowe lasso [5], które wybiera podzbiór zmiennych ilościowych i podzbiór zmiennych jakościowych. Metoda ta albo usuwa albo nie całe czynniki. W grupie tej znajduje się także algorytm używający niewypukłej regularyzacji MCP [2], która jest połączeniem kary l_1 (lasso) i l_0 (liczba parametrów w modelu). Do drugiej grupy należy algorytm CAS-ANOVA [1], który wybiera podzbiór zmiennych ciągłych oraz podziały zbiorów wartości dla czynników. Metoda ta skleja poziomy zmiennych jakościowych. Efektywniejszą implementacją algorytmu CAS-ANOVA jest algorytm gvcm [4]. Wszystkie te metody używają regularyzacji.

W pracy doktorskiej jest opisany nowy algorytm o nazwie DMR (Delete or Merge Regressors, [3]). Tak jak CAS-ANOVA wybiera on podzbiór zmiennych ilościowych i podziały dla zmiennych jakościowych. Jednakże, zamiast regularyzacji używa zachłannej wersji wyboru podzbioru predyktorów. Najpierw tworzona jest zagnieżdżona rodzina modeli różniących się od siebie o jeden parametr, poprzez usuwanie zmiennej ciągłej albo sklejanie dwóch poziomów czynnika, gdzie kolejność akceptowania kolejnych hipotez jest oparta o sortowanie statystyk testu ilorazu wiarogodności. Następnie ostateczny model jest wybierany za pomocą kryterium informacyjnego.

Algorytm DMR działa tylko dla danych, gdzie $p < n$ (liczba kolumn w macierzy planu jest mniejsza niż liczba obserwacji). W pracy doktorskiej zaproponowane jest rozszerzenie DMR czyli algorytm DMRnet, który działa także dla problemów gdzie $p \gg n$. DMRnet używa regularyzacji w kroku przesiewu predyktorów za pomocą grupowego lasso, a następnie wykonuje procedurę DMR po zmniejszeniu macierzy planu do wymiaru $p < n$.

Praktyczne wyniki dotyczą sześciu zbiorów danych rzeczywistych i dwunastu scenariuszy symulacyjnych. Pokazano, że DMRnet wybiera mniejsze modele przy prawie minimalnym błędzie predykcji w stusunku do konkurencyjnych metod. Ponadto, w scenariuszach symulacyjnych DMRnet najczęściej wybiera prawdziwe modele.

Teoretyczne wyniki pracy to twierdzenia, że algorytmy DMR dla regresji liniowej i logistycznej wybierają model prawdziwy z prawdopodobieństwem rosnącym do jednego nawet jeśli p dąży do nieskończoności wraz z n. Ponadto, podane są górne ograniczenia na błąd selekcji.

Literatura

[1] Bondell, Howard D., and Brian J. Reich. *Simultaneous factor selection and collapsing levels in ANOVA.* Biometrics 65.1 (2009): 169-177.

[2] Breheny, Patrick, and Jian Huang. *Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors.* Statistics and computing 25.2 (2015): 173-187.

[3] Maj-Kańska, Aleksandra, Piotr Pokarowski, and Agnieszka Prochenka. *Delete or merge regressors for linear model selection.* Electronic Journal of Statistics 9.2 (2015): 1749-1778.

[4] Oelker, Margret-Ruth, Jan Gertheiss, and Gerhard Tutz. *Regularization and model selection with categorical predictors and effect modifiers in generalized linear models.* Statistical Modelling 14.2 (2014): 157-177.

[5] Yuan, Ming, and Yi Lin. *Model selection and estimation in regression with grouped variables.* Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68.1 (2006): 49-67.