Статистические методы отбора значимых переменных

А.В. Булинский

(Механико-математический факультет
МГУ им. М.В.Ломоносова)

Ломоносовские чтения
МГУ, 26 декабря 2014 года

ПЛАН

ПЛАН

1. Введение
2. Критерий сильной состоятельности оценок функционала ошибки прогноза случайного отклика
3. Центральная предельная теорема для регуляризованных статистик
4. Результаты компьютерного моделирования

ПЛАН

1. Введение
2. Критерий сильной состоятельности оценок функционала ошибки прогноза случайного отклика
3. Центральная предельная теорема для регуляризованных статистик
4. Результаты компьютерного моделирования

ПЛАН

1. Введение
2. Критерий сильной состоятельности оценок функционала ошибки прогноза случайного отклика
3. Центральная предельная теорема для регуляризованных статистик
4. Результаты компьютерного моделирования

ПЛАН

1. Введение
2. Критерий сильной состоятельности оценок функционала ошибки прогноза случайного отклика
3. Центральная предельная теорема для регуляризованных статистик
4. Результаты компьютерного моделирования

ПЛАН

1. Введение
2. Критерий сильной состоятельности оценок функционала ошибки прогноза случайного отклика
3. Центральная предельная теорема для регуляризованных статистик
4. Результаты компьютерного моделирования

Выявлению значимых факторов, описывающих бинарный случайный отклик, посвящена обширная литература. В медико-биологических исследованиях имеется специальная область GWAS (genome-wide association studies), в которой осуществляется анализ связей фенотипа и генотипа. В частности, определяются факторы, повышающие риск сложных заболеваний таких как гипертензия, диабет, инфаркт миокарда.

Среди ряда взаимодополняющих средств статистического анализа генетических данных отметим современные варианты метода главных компонент (S.Lee et al., 2012), логистическую и логическую регрессию (K.Sikorska et al., 2013; H.Schwender, I.Ruczinski, 2010), LASSO (R.Lockhart et al., 2013; R.J.Tibshirani, J.Taylor, 2012), байесовскую технику и разнообразные методы машинного обучения (H.Hastie et al., 2008).

Подчеркнем, что существуют различные модификации упомянутых методов, см., например,

T.Koski, J.M.Noble, Wiley, 2009;

J.C.Dunlap, J.H.Moore, Academic Press, 2010,

R.Jiang et al. (eds.), Springer, 2013.

В докладе проводится сравнение основных статистических методов понижения размерности факторов $X_1, \ldots, X_n$, описывающих некоторый случайный отклик Y, с предложенным новым методом отбора значимых факторов $X_{i_1}, \ldots, X_{i_r}$, где $\{i_1, \ldots, i_r\} \subset \{1, \ldots, n\}$.

Развиты недавние исследования, начатые в МГУ под руководством академика В.А.Садовничего (A.Bulinski, O.Butkovsky, V.Sadovnichy, A.Shashkin, P.Yaskov, A.Balatskiy, L.Samokhodskaya, V.Tkachuk Statistical methods of SNP data analysis and applications. Open Journal of Statistics, 2012, volume 2, No 1, p. 73-87).

В 2014 году удалось продвинуться в области развития MDR (multifactor dimensionality reduction) метода, т.е. метода понижения размерности факторов. Этот метод был предложен в известной статье M.Ritchie et al. (2001), посвященной бинарному отклику и построению для соответствующих наблюдений зон высокого и низкого риска.

В последующие годы появилось более 200 работ, развивающих этот метод и содержащих его применения.

В докладе излагаются результаты следующего цикла работ.

[1] А.В. Булинский. К основам метода понижения размерности объясняющих переменных. Записки научных семинаров ПОМИ им. В.А.Стеклова, 2012. том 408, с. 84-101.

[2] A.Bulinski. Central limit theorem related to MDR method. Proceedings of the Fields Institute. Int. Workshop in Honour of Professor Csorgo's work. p. 1-15 (принята к печати).

[3] А.В.Булинский и А.С.Ракитько. Оценивание бинарного отклика. Доклады Академии наук. 2014, т. 455, N 6, с. 1-5.

[4] A.Bulinski. Some statistical methods in genetics. Lecture Notes in Mathematics, v. 2120, p. 293-320, Springer, 2014.

[5] A.V.Bulinski, A.S.Rakitko MDR method for nonbinary response variable. Journal of Multivariate Analysis (принята к печати).

[6] A.V.Bulinski, A.S.Rakitko Simulation and analytical approach to the identification of significant factors. Communications in Statistics - Simulation and Computation (принята к печати).

Подчеркнем, что предложен новый подход к идентификации значимых факторов, влияющих на небинарный случайный отклик. Для этого введены статистические оценки функционала ошибки прогноза отклика, вовлекающие штрафную функцию (или ее оценки) и процедуру кросс-валидации. Установлен критерий сильной состоятельности оценок введенного функционала.

Это привело к обоснованному и унифицированному выявлению значимых факторов (влияющих на изучаемый случайный отклик). Более того, при весьма широких условиях для регуляризованных версий этих оценок доказана центральная предельная теорема, позволившая характеризовать устойчивость статистических выводов.

Самостоятельный интерес представляет новый вариант центральной предельной теоремы, полученной для массивов перестановочных случайных величин. Осуществлено также компьютерное моделирование, которое показало эффективность предложенного подхода.