Секция «Теория вероятностей и математическая статистика»

Последовательный отбор переменных в MDR-EFE методе.

Научный руководитель – Булинский Александр Вадимович

Ракитько Александр Сергеевич
Аспирант
Московский государственный университет имени М.В.Ломоносова,
Механико-математический факультет, Кафедра теории вероятностей, Москва, Россия
E-mail: rakitko@gmail.com

Задача выявления факторов, объясняющих некоторый случайный отклик Y, возникает во многих прикладных исследованиях. Например, в медико-биологических исследованиях в качестве факторов могут выступать генетические маркеры (SNP) $X_1, \dots, X_n$, принимающие дискретные значения, а Y показывает наличие или отсутствие заболевания (соответственно, значения 1 или -1). Как правило, число исследуемых факторов n много больше количества имеющихся N наблюдений. При этом считается, что количество значимых факторов $X_{k_1}, \dots, X_{k_r}$, влияющих на Y, невелико. Для поиска таких факторов применяется различная техника (LARS, LASSO, MDR, Bayes analysis и другие).

Нами используется метод MDR-EFE (Multifactorial Dimensionality Reduction with Error Function Estimation) понижения размерности набора факторов, развитый в [1,2]. Этот метод основан на анализе статистической оценки функционала ошибки предсказания отклика. Данный функционал задается формулой $Err(f_{PA}) = |Y - f_{PA}(X)|\psi(Y)$, где f_{PA} – предсказательный алгоритм, а $\psi(\cdot)$ – штрафная функция. В указанных работах [1,2] был установлен критерий сильной состоятельности предлагаемых оценок функционала ошибки, а также доказаны варианты центральной предельной теоремы для введенных статистик. Метод применим к широкому классу моделей, но имеет высокую вычислительную сложность, поскольку приходится перебирать большое число комбинаций факторов.

Для упрощения и ускорения алгоритма предлагается использовать последовательный отбор переменных (forward selection), см., например, [3]. А именно, на первом шаге выбирается фактор X_{i_1}, имеющий наименьшую оценку функционала ошибки. На втором шаге выбирается фактор X_{i_2}, для которого пара (X_{i_1}, X_{i_2}) имеет наименьшую оценку ошибки среди пар вида $(X_{i_1}, \cdot)$, и так далее.

Будем предполагать, что справедлива модель наивного байесовского классификатора, т.е.

$$\mathsf{P}(X_1 = x_1, \dots, X_n = x_n | Y = y) = \prod_{i=1}^{n} \mathsf{P}(X_i = x_i | Y = y)$$

для всех возможных значений $x_1, \dots, x_n, y$. Такое предположение приводит к некоторой модели логистической регрессии и позволяет дать нижнюю оценку вероятности события A, состоящего в правильном нахождении r значимых факторов при описанном выше алгоритме последовательного поиска.

Теорема. *Пусть выполнены условия, обеспечивающие сильную состоятельность и асимптотической нормальности оценки функционала ошибки предсказания отклика (см. [1,2]). Тогда*

$$\mathsf{P}(A) \geq \prod_{k=1}^{n-2} \left(1 - \frac{1}{N} \sum_{t=k+2}^{n} \frac{V_{max} + o(1)}{\left(c_{k+1,t}^{(1,\dots,k)} + o(1) \right)^2} \right), \quad N \to \infty,$$

где V_{max} – некоторая константа, а величины $c_{k+1,t}^{(1,\dots,k)}$ определенным образом выражаются через коэффициенты логистической регрессии.

С помощью компьютерного моделирования проведена оценка качества работы указанного алгоритма для разных конфигураций модели наивного байесовского классификатора. Кроме того, рассмотрен вариант алгоритма с регуляризованной версией функционала ошибки.

Источники и литература

1) Alexander Bulinski and Alexander Rakitko. MDR method for nonbinary response variable. *Journal of Multivariate Analysis*, 135:25 – 42, 2015.

2) Alexander Bulinski and Alexander Rakitko. Simulation and analytical approach to the identification of significant factors. *Communications in Statistics - Simulation and Computation*, 45(5):1430–1450, 2016.

3) Alan Jović, Karla Brkić, and Nikola Bogunović. A review of feature selection methods with applications. *Information and Communication Technology, Electronics and Microelectronics* (MIPRO), 2015 38th International Convention on. IEEE, 2015.