

РЕГРЕССИОННЫЕ МОДЕЛИ QSAR

Бердник М. И., Дяченко А. В., Иванов В. В.

Харьковский национальный университет имени В. Н. Каразина

berdnikm@i.ua

На сегодняшний день задача получения регрессионных уравнений типа «структура-свойство» становится актуальной во многих отраслях химии. Регрессионные модели незаменимы и при разработке новых лекарственных соединений и при поиске материалов с заданными свойствами. При этом основная масса расчетов проводится с использованием стандартного метода наименьших квадратов (*Ordinary Least Squares, OLS*). Однако, на сегодняшний день, в научной литературе обсуждается также и ряд альтернативных подходов к построению регрессионных уравнений. Среди них метод наименьших модулей (*Least Absolute Deviation, LAD*), регрессия на главных компонентах (*Principle Component Regression, PCR*), неполный метод наименьших квадратов (*Partial Least Squares, PLS*), метод **LASSO** (*Least Absolute Selection and Shrinkage Operator*), метод наименьших углов (*Least Angle Regression Stagewise, LARS*), метод ортогональных расстояний (*Orthogonal Distance Regression, ODR*) и т. д. Вместе с тем, не смотря на многообразие методов, в литературе практически не представлены работы, в которых бы обсуждалась *сравнительная* эффективность и применимость этих подходов. Целью настоящей работы было изучение данных вопросов в применении к конкретным проблемам поиска количественных соотношений структура-свойство в химии. Для этого нами разработан комплекс компьютерных программ реализующий вышеперечисленные методы. Рассматривались проблемы построения регрессионных уравнений для описания pK_a , температур кипения, давления насыщенного пара и т. д. органических соединений различных классов.

Для иллюстрации приведем некоторые результаты, касающиеся описания pK_a для простой тестовой выборки пятнадцати органических кислот. На основании методов **LASSO** и **LARS** из более чем двух тысяч дескрипторов было отобрано три: заряд на кислороде карбонильной группы (x_1), оксигруппы (x_2) и на водороде оксигруппы (x_3). По результатам **OLS** и **LAD** расчетов были получены следующие уравнения. Однопараметрические:

$$\text{OLS: } pK_a = -24.44 - 91.35x_2, \quad R^2 = 0.852, \quad Q^2 = 0.805, \quad (1)$$

$$\text{LAD: } pK_a = -20.53 - 79.06x_2, \quad R^2 = 0.839, \quad Q^2 = 0.798. \quad (2)$$

Трехпараметрические:

$$\text{OLS: } pK_a = -1.08 - 19.93x_1 - 46.80x_2 - 67.61x_3, \quad R^2 = 0.971, \quad Q^2 = 0.746, \quad (3)$$

$$\text{LAD: } pK_a = -4.28 - 17.22x_1 - 55.12x_2 - 61.17x_3, \quad R^2 = 0.969, \quad Q^2 = 0.201. \quad (4)$$

Здесь R^2 и Q^2 коэффициенты детерминации. Причем Q^2 соответствует процедуре перекрестного оценивания (*leave-one-out cross validation, LOO-CV*). Как видим, оба однопараметрических уравнения полученных в **OLS** и **LAD** практически одинаковы в прогностической способности. Трехпараметрическое уравнение **OLS** достаточно точно описывает искомое свойство, в то время как уравнение **LAD** оказывается неадекватным, $R^2 - Q^2 \approx 0.8$. Применение метода **PCR** показало, что использование двух сингулярных чисел является оптимальным для данной задачи $R^2 \approx Q^2 \approx 0.95$.