Статистические методы в QSAR

Часть 1

Множественная линейная регрессия

Затронутые темы

- Задача статистического анализа в QSAR
- Множественная линейная регрессия
- Статистические показатели: R, s, F
- Пошаговый вариант множественной линейной регрессии
- Разбиение выборки на обучающую и контрольную
- Процедура скользящего контроля

Задача статистического анализа в QSAR

Целью статистического анализа в QSAR является поиск функции *f*, связывающей значение свойства *y* (которое может быть как физико-химическим свойством, так и биологической активностью) со значениями дескрипторов *x*1,…,*xM*, описывающих химическое соединение:

$$y \propto f(x_1,...,x_M)$$

Y непрерывное – регрессионный анализ, аппроксимация функции

Y дискретное – дискриминантный анализ, классификация, распознавание образов

Задача статистического анализа в QSAR

$$y^j = f(x_1^j, ..., x_M^j) + e^j$$

j=1,…,N

$$y^j = F(c_1, ..., c_P; x_1^j, ..., x_M^j) + e^j$$

Принцип максимального правдоподобия

$$L(c_1, ..., c_P) \to \max$$

Метод наименьших квадратов

- Выборка является репрезентативной
- Случайная величина ε имеет нормальное распределение
- Наблюдения являются независимыми
- Наблюдения являются равноточными

$$\sum_{j=1}^{N} (e^j)^2 \to \min$$

Множественная линейная регрессия – постановка задачи

$$y \propto c_0 + c_1 x_1 + \ldots + c_M x_M$$

$$y^j = c_0 + \sum_{i=1}^{M} c_i x_i^j + e^j \qquad j = 1,\ldots,N$$

Найти такие значения C_i: $\sum_{j=1}^{N} (e^j)^2 \to \min$

Множественная линейная регрессия – решение задачи

$$C = (X^T X)^{-1} X^T Y$$

$$C = \begin{pmatrix} c_0 \\ c_1 \\ \vdots \\ c_M \end{pmatrix}$$

Регрессионные коэффициенты

$$Y = \begin{pmatrix} y^1 \\ y^2 \\ \vdots \\ y^N \end{pmatrix}$$

Экспериментальные значения свойства

$$X = \begin{pmatrix} 1 & x_1^1 & \cdots & x_M^1 \\ 1 & x_1^2 & \cdots & x_M^2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_1^N & \cdots & x_M^N \end{pmatrix}$$

Значения дескрипторов

Статистические показатели для МЛР

RSS – сумма квадратов остатков

$$RSS = \sum_{j=1}^{N} (e^j)^2$$

SS – дисперсия свойства Y

$$SS = \sum_{j=1}^{N} (y^j - \bar{y})^2 \qquad \bar{y} = \frac{1}{N} \sum_{j=1}^{N} y^j$$

R – коэффициент корреляции

$$R = \sqrt{\frac{SS - RSS}{SS}}$$

O<R<1

Radj – скорректированный коэффициент корреляции

$$R_{adj} = \sqrt{R^2 - (1 - R^2)\frac{M}{N - M - 1}}$$

Статистические показатели для МЛР

RMSE – среднеквадратичное значение ошибки

$$RMSE = \sqrt{\frac{RSS}{N}}$$

s – стандартное отклонение

$$s = \sqrt{\frac{RSS}{N - M - 1}}$$

MAE – средняя абсолютная ошибка

$$MAE = \frac{1}{N}\sum_{j=1}^{N}\left|e^{j}\right|$$

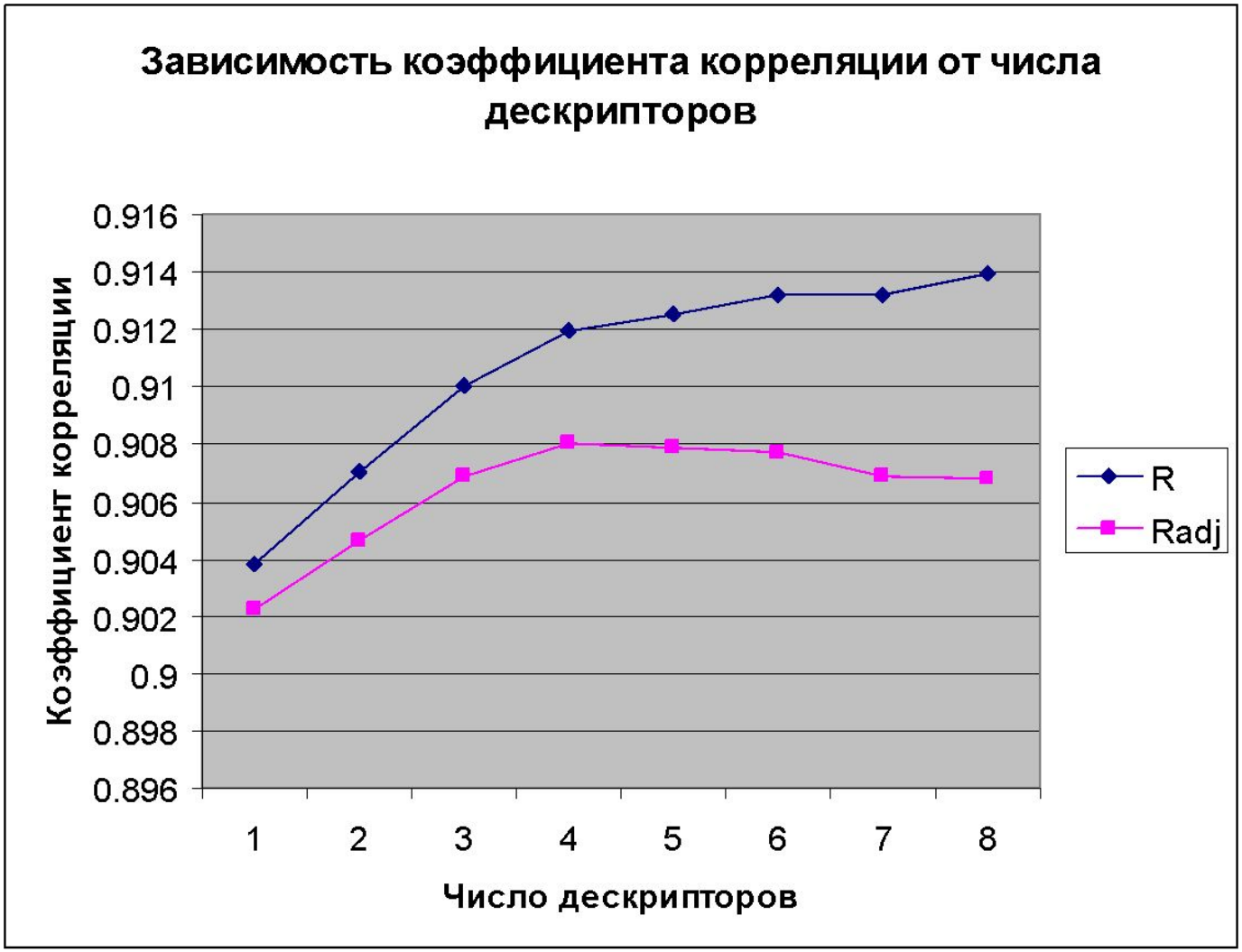
F – критерий Фишера

$$F = \frac{(SS - RSS)/(M + 1)}{RSS/(N - M - 1)}$$

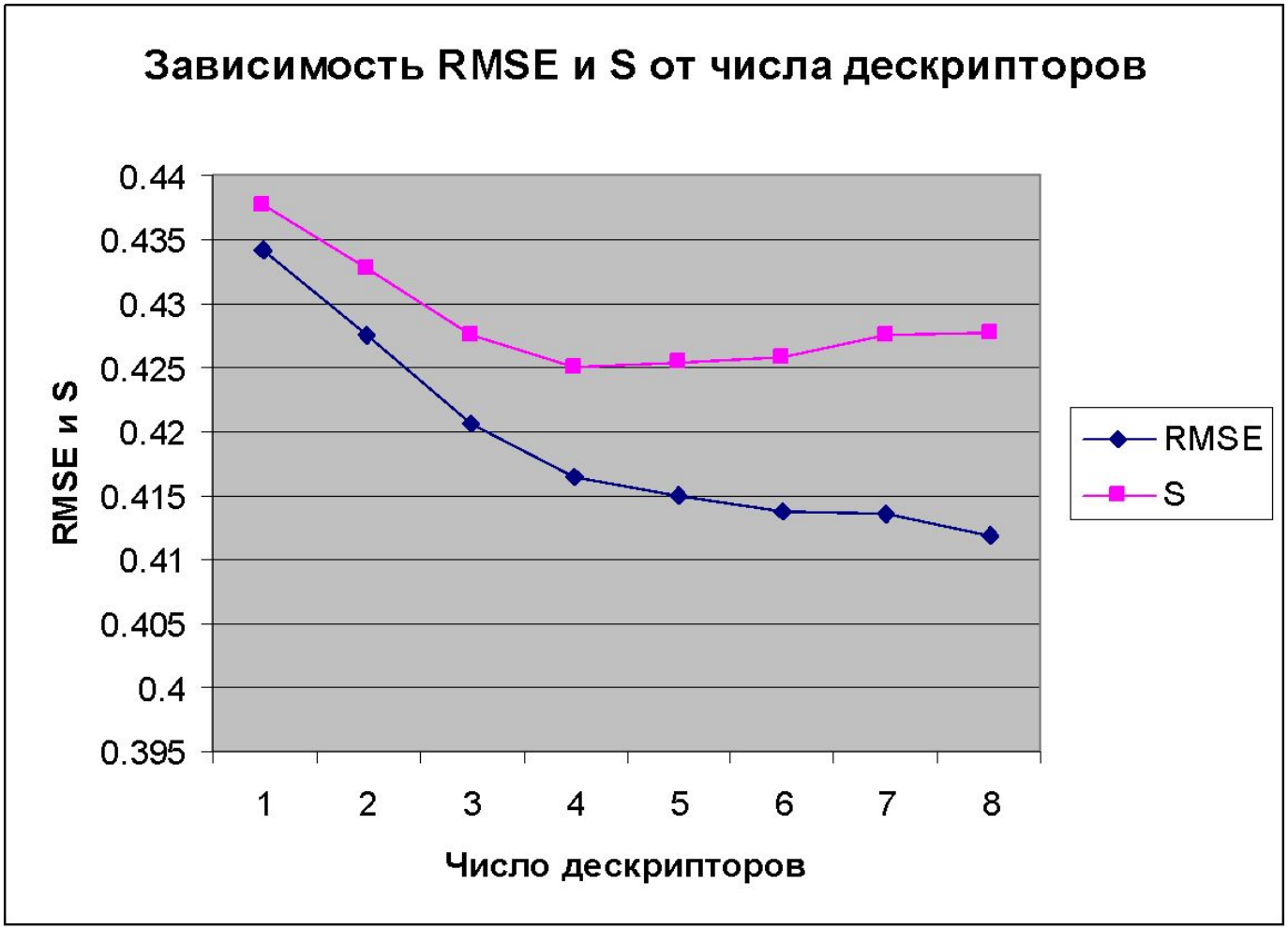
Статистические показатели для МЛР

Показатели описательной способности линейной регрессионной модели	Показатели прогнозирующей способности линейной регрессионной модели
R	R_{adj}
$RMSE$	s
MAE	
	F

Статистические показатели для МЛР



Статистические показатели для МЛР



Статистические показатели для МЛР

Индивидуальный t-критерий (критерий Стьюдента) для дескриптора

$$t_i = \frac{c_i}{s \cdot \sqrt{a_{ii}}}$$

$$\left\| a_{ij} \right\| = (X^T X)^{-1}$$

Необходимость отбора дескрипторов

- Проблема мультиколлинеарности дескрипторов и сингулярности матрицы $(X^TX)^{-1}$
- Проблема переопределенности моделей
- Внесение «шума» в модель нерелевантными дескрипторами

Пошаговый вариант множественной линейной регрессии

$$F_{ie} = \frac{(R_2^2 - R_1^2) \cdot (N - M_2 - 1)}{1 - R_2^2}$$

Дескриптор включается в модель, если $F_{ie} = \max \quad \wedge \quad F_{ie} > F_{IN}$

Дескриптор исключается из модели, если $F_{ie} = \min \quad \wedge \quad F_{ie} < F_{OUT}$

Типичные значения порогов: F_{IN} = 3.84, F_{OUT} = 2.7

Разбиение выборки на обучающую и контрольную

***PRSS**s* - сумма квадратов остатков при прогнозе $$PRSS_S = \sum_{j \in S} (e^j)^2$$

***PSS**s* - дисперсия свойства *y* на контрольной выборке $$PSS_S = \sum_{j \in S} (y^j - \bar{y})^2$$

***PRMSE**s* - среднеквадратичная ошибка на прогнозе $$PRMSE_S = \sqrt{\frac{PRSS_S}{N_S}}$$

***PMAE**s* - средняя абсолютная ошибка на прогнозе $$PMAE_S = \frac{1}{N_S} \sum_{j \in S} \left| e^j \right|,$$

Q^2_s - квадрат коэффициента корреляции на прогнозе $$Q^2_S = \frac{PSS_S - PRSS_S}{PSS_S}$$

Процедура скользящего контроля (cross-validation)

1. При μ-кратном скользящем контроле исходная выборка разбивается на μ приблизительно равных частей
2. Каждая из этих частей по очереди объявляется контрольной выборкой
3. Для нее формируется обучающая выборка, состоящая из всех соединений из исходной выборки, в нее не входящих
4. По обучающей выборке строится регрессионная модель
5. По текущей контрольной выборке вычисляется сумма квадратов ошибок PRSSs и сумма абсолютных ошибок PSAEs
6. Пункты 2-5 повторяются для всех μ частей

Процедура скользящего контроля – статистические показатели

$$PRSS_{CV} = \sum_{k=1}^{\mu} PRSS_{S_k}$$

$$PSAE_{CV} = \sum_{k=1}^{\mu} PSAE_{S_k}$$

***RMSE*$_{CV}$** - среднеквадратичная ошибка прогноза

$$RMSE_{CV} = \sqrt{\frac{PRSS_{CV}}{N}}$$

***MAE*$_{CV}$** - средняя абсолютная ошибка прогноза

$$MAE_{CV} = \frac{PSAE_{CV}}{N}$$

***Q*$^2_{CV}$** – коэффициент корреляции для прогноза

$$Q^2_{CV} = \frac{SS - PRSS_{CV}}{SS}$$