Krymova E.A., Strijov V.V. Model selection and multicollinearity analysis // Proceedings of conference on Intelligent data processing, 2010 : 153-156.

Выбор регрессионных моделей с анализом мультиколлинеарности*

Крымова Е. А., Стрижов В. В.
ekkrym@gmail.com
Москва, Вычислительный центр РАН

В работе рассматриваются способы порождения моделей с помощью существенно нелинейных параметрических порождающих функций. Предложен алгоритм выбора модели оптимальной структуры, основанный на последовательном порождении моделей максимального правдоподобия. Исследуется расстояние между полученными моделями. Работа алгоритма выбора моделей проиллюстрирована задачей моделирования давления в двигателе внутреннего сгорания.

Feature selection with multicollinearity control in regression analysis*

Krymova K. A., Strijov V. V.
Moscow, Russian Academy of Sciences, Computing Center

We investigate the regression model construction problem. Assume the given set of features to be inefficient to construct the adequate model. We generate new features on the base of a given sample set. The number of the features exceeds the number of the objects after the generation. The features require reduction of multicollinearity. The new feature selection algorithm is developed. Coherent Bayesian Inference is used for feature selection: the set of indices is necessary which corresponds the most evident model. Modelling of the pressure in diesel engine is used as a practical example.

Рассматривается задача выбора признаков при построении линейной регрессионной модели. Предполагается, что исходных признаков недостаточно для построения адекватной регрессионной модели. Предлагается породить новый набор признаков, модифицируя исходные. Будем считать, что для получения адекватной модели необходимо сделать выбор из большого числа признаков.

В работе описан алгоритм отбора признаков, который заключается в последовательном добавлении и удалении признаков. При выборе признаков используется принцип максимума правдоподобия: отыскивается набор признаков, соответствующий модели с максимальным значением правдоподобия [1, 2]. Оцениваются параметры и правдоподобие самых моделей. Шаги добавления и удаления проводятся таким образом, что с увеличением номера шага правдоподобие моделей возрастает.

После добавления признаков возникает проблема мультиколлинеарности, когда имеются признаки, сильно коррелирующие друг с другом. Существуют следующие способы обнаружения мультиколлинеарности: проверка корреляции между признаками [3], исследование факторов инфляции дисперсии (VIF) [4], метод Белсли [5]. Описываемый алгоритм удаляет признаки согласно методу Белсли. Результатом работы алгоритма является наиболее правдоподобная модель, включающая наименее коррелирующие признаки.

Результаты работы предложенного алгоритма сравниваются с результатами работы алгоритма LARS [6]. Сравнение проводится на задаче моделирования давления в камере двигателя внутреннего сгорания.

Постановка задачи

Задана выборка $D = \{(\mathbf{x}^i, y^i)\}_{i=1}^m$ — множество m пар, $\mathbf{x}^i = (x^i_j)_{j=1}^n$, $\mathbf{x}^i \in \mathbb{R}^n$ — вектор значений n признаков, и одной зависимой переменной $y^i \in \mathbb{R}^1$. Индекс i элементов выборки и индекс j признака далее будем рассматривать как элементы множеств $i \in I = \{1, \dots, m\}$ и $j \in J = \{1, \dots, n\}$.

Задан класс регрессионных моделей $\mathcal{F} = \{f_s\}$ — параметрических функций, линейных относительно параметров,

$$y^i = f_s(\mathbf{w}_s, \mathbf{x}^i) = \sum_{j \in J_s} w_j x^i_j, \qquad (1)$$

в которой $s \in \{1, \dots, 2^n\}$ является индексом модели, $\mathbf{w}_s = (w_j)_{j \in J_s}$ — вектор параметров, заданный индексом модели, $J_s \subseteq J$ — набор индексов признаков. Введено ограничение на число элементов линейной комбинации (1). В множество $\mathcal{F}$ могут входить только модели с числом признаков $|J_s| \leqslant R$.

Принята следующая гипотеза порождения данных. Пусть случайная аддитивная переменная ν регрессионной модели

$$y = f(\mathbf{w}, \mathbf{x}) + \nu$$

имеет нормальное распределение $\mathcal{N}(0, \sigma^2_\nu)$.

Тогда, с учетом гомоскедастичности регрессионных остатков, распределение зависимой переменной имеет вид

$$p(y|x, \mathbf{w}, \sigma^2_\nu, f) = \frac{\exp(-\frac{1}{\sigma^2_\nu} S(D|\mathbf{w}, f))}{(2\pi\sigma^2_\nu)^{\frac{n}{2}}}, \qquad (2)$$

Работа выполнена при финансовой поддержке РФФИ, проект №10-07-00422-а.

где S — сумма квадратов невязок $y^i - f(\mathbf{w}, \mathbf{x}^i)$. Это распределение задает указанный ниже критерий качества модели.

Дополнительно задано разбиение выборки $I = I^{\mathrm{T}} \sqcup I^{\mathrm{C}}$ на обучающую и контрольную. Для каждого набора данных, рассматриваемого в вычислительном эксперименте, наборы индексов $I^{\mathrm{T}}, I^{\mathrm{C}}$ определены до начала эксперимента. Алгоритм выбора модели определяет метод оптимизации, доставляющий оптимальное значение параметрам $\mathbf{w}$ модели f на обучающей выборке $\{(x^i, y^i)\colon i \in I^{\mathrm{T}}\}$. Принят критерий качества — сумма квадратов регрессионных остатков на контрольной выборке

$$S = \sum_{i \in I^{\mathrm{C}}} \left(y^i - f(\mathbf{w}, \mathbf{x}^i)\right)^2. \qquad (3)$$

Требуется найти такую модель $f_s \in \mathcal{F}$, которая доставляет наименьшее значение функционалу качества (3). Такая модель будет называться моделью оптимальной структуры.

Порождение признаков

Предлагается следующий способ формирования выборки D.

Задано множество признаков $\Xi = \{\xi^u\}_{u=1}^{U}$ и конечное множество функций $G = \{g_v\}_{v=1}^{V}$. Рассмотрим декартово произведение $G \times \Xi$, элементу (g_v, ξ^u) которого поставлена в соответствие суперпозиция $g_v(\xi^u)$, однозначно определяемая индексами v, u. Обозначим $a_i = g_v(\xi^u)$, где индекс $i = (v-1)U + u$.

Назначается базовая модель порождения признаков. В качестве модели, описывающей отношение между зависимой переменной y и переменными a_i, используется полином:

$$y = w_0 + \sum_{\iota=1}^{UV} w_\iota a_\iota + \sum_{\iota=1}^{UV}\sum_{\zeta=1}^{UV} w_{\iota\zeta} a_\iota a_\zeta + \ldots$$
$$\cdots + \sum_{\iota_1=1}^{UV} \cdots \sum_{\iota_z=1}^{UV} w_{\iota_1 \ldots, \iota_z} a_\iota \ldots a_{\iota_z},$$

где вектор коэффициентов

$$\mathbf{w} = (w_0, w_\iota, w_{\iota\zeta}, \ldots, w_{\iota_1 \ldots \iota_z})_{\iota, \zeta, \ldots, \iota_1, \ldots \iota_z = 1, \ldots, m}.$$

Запишем вышеприведенный ряд в виде

$$y = \sum_{j \in J} w_j x^j.$$

Переменные $\{x^j\}$ поставлены в однозначное соответствие мономам полинома.

Алгоритм максимизации правдоподобия

В качестве исходного алгоритма, с которым будет производиться сравнение, был выбран алгоритм LARS, описанный в [6].

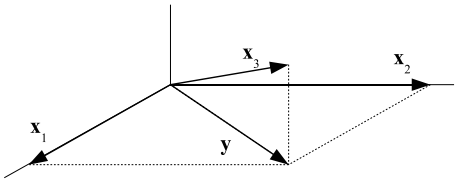


Рис. 1. Пример, иллюстрирующий последовательность выбора признаков.

Для иллюстрации основного недостатка алгоритма LARS рассмотрим следующий пример 1. Пусть матрица X состоит столбцов значений трех признаков. Первый признак $\mathbf{x}_3$ сильно коррелирует с вектором ответов $\mathbf{y}$, который является линейной комбинацией остальных двух признаков $\mathbf{x}_1$ и $\mathbf{x}_2$. LARS на первом шаге выберет признак $\mathbf{x}_3$, так как он сильнее коррелирует с вектором ответов и затем присоединит остальные признаки. Для разрешения этого недостатка предложен алгоритм, позволяющий удалять мультиколинеарные признаки и добавлять признаки, уменьшаюшие ошибку.

Используем два набора признаков: набор $\mathcal{Z}$, который содержит все признаки, и текущий набор $\mathcal{A}$. В начале работы алгоритма $\mathcal{A} = \varnothing$. Рассмотрим k-й шаг алгоритма.

1. Последовательно, из набора $\mathcal{Z} \setminus \mathcal{A}_k$ в текущий набор признаков $\mathcal{A}_k$ добавляются признаки, наиболее коррелирующие с вектором регрессионных остатков.
2. Выполняется прореживание модели: последовательно удаляются те элементы линейной комбинации, заданной набором $\mathcal{A}_k$, для которых критерий мультиколлинеарности принимает максимальное значение.

Добавление (удаление) признаков происходит при возрастающем правдоподобии при недостаточном (избыточном) числе признаков в текущем наборе; когда в наборе окажется избыточное (недостаточное) число признаков, правдоподобие начнет уменьшаться. Шаги добавления или удаления продолжаются до тех пор, пока, число последовательных шагов при уменьшающемся правдоподобии не превзойдет заданное число K.

Сравнение моделей выполняется с помощью связанного Байесовского вывода. Пусть многомерная случайная величина — вектор параметров модели имеет нормальное распределение с нулевым математическим ожиданием и дисперсией α^{-2}. Тогда распределение вектора параметров модели

$$p(\mathbf{w}|\alpha, f) = \frac{1}{(2\pi/\alpha)^{\frac{n}{2}}} \exp\left(-\frac{\alpha}{2}\mathbf{w}^{\intercal}\mathbf{w}\right). \qquad (4)$$

Пусть в (2) $\beta = \sigma_\nu^{-2}$. Тогда α и β — гиперпараметры модели.

Правдоподобие $p(D|f)$ модели f при фиксированных гиперпараметрах:

$$p(D|f,\alpha,\beta) = \int p(D|f,\mathbf{w},\beta,\alpha)p(\mathbf{w}|f,\alpha)d\mathbf{w}. \quad (5)$$

Для каждой модели f, заданной множеством индексов признаков, вычисляются наиболее правдоподобные параметры и гиперпараметры модели.

Обозначим распределение моделей при фиксированных данных $p(f_i|D)$ и рассмотрим числитель формулы Байеса

$$p(f_i|D) = \frac{p(D|f_i)p(f_i)}{p(D)}, \quad (6)$$

в котором правдоподобие моделей $p(D|f_i)$ определяется выражением (5). Будем считать априорную вероятность равной для всех моделей, $p(f_i) = \text{const}$. Так как знаменатель выражения (6) не зависит от выбора модели, то сравнение моделей происходит через вычисление правдоподобия моделей.

Результатом работы алгоритма является модель удовлетворительной точности; мультикоррелирующие признаки исключены.

Критерий мультиколлинеарности

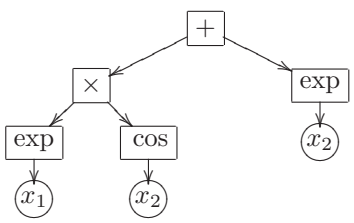
Предложенный алгоритм при исследовании мультиколлинеарости использует метод Белсли.

Проводится сингулярное разложение [7] матрицы признаков $X = U\Lambda V^\intercal$, где $V = (v_{ij})$ — ортогональная матрица, столбцы которой являются собственными числами матрицы ковариаций $X^\intercal X$. Индекс обусловленности с номером i — отношение максимального сингулярного числа к i-му сингулярному числу $\eta_i = \lambda_{\max}/\lambda_i$.

Наличие больших индексов обусловленности η_j означает что есть зависимость между признаками. Рассматриваются собственные векторы матрицы ковариаций соответствующие большим индексам обусловленности. Можно показать, что большие значения $\left(v_{ij}^2/\eta_j^2\right)$ соответствуют коррелирующим признакам.

Расстояние между моделями

Для исследования сходимости введем расстояние между моделями. Модель представлена в виде дерева: нелистовыми вершинами которого являются функции, листовые —признаки. Например, модель $y = w_1 \exp x_1 \cos x_2 + w_2 \exp x_2$ имеет следующее дерево:



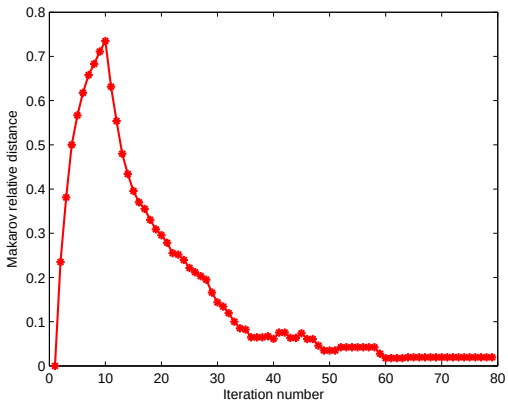


Рис. 2. Зависимость максимального расстояния Макарова R_{ij} между моделями из множества 10 последовательно порожденных моделей от номера итерации.

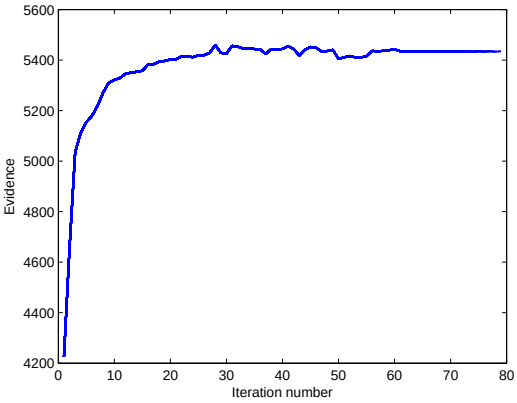


Рис. 3. График значений правдоподобия в зависимости от номера итерации.

Пусть $H_i(V,Z)$ — конечный, помеченный, связный, неориентированный граф без петель и кратных ребер, имеющий множество вершин V_i, $|V_i| = p_i > 0$, и множество ребер Z_i, $|Z_i| = q_i > 0$.

В [8] предложено использовать меру структурного подобия графов, зависящую от числа вершин или ребер их наибольшего общего подграфа, и предложены некоторые меры. Наибольший общий подграф графов H_i и H_j обозначим через $H_{ij}(V_{ij},Z_{ij})$, $|V_i|=p_i>0$, $|Z_i|=q_i>0$, $|V_j|=p_j>0$, $|Z_j| = q_j > 0$, $|V_{ij}| = p_{ij} \geqslant 0$, $|Z_{ij}| = q_{ij} \geqslant 0$.

Предложены следующие функция расстояния (и показано, что они являются метриками):

$$r_{ij} = q_i + q_j - 2q_{ij};$$
$$R_{ij} = r_{ij}/(q_i + q_j).$$

Вычислительный эксперимент

Работа алгоритма проиллюстрирована на примере данных измерения давления в двигателе внутреннего сгорания. Выборка состоит из набора временных рядов — измерений давления в камере

Таблица 1. Результаты работы алгоритмов выбора признаков.

Алгоритм	CV	AIC	$\lg \varkappa$	k
LARS	0,011	−1346	7	49
Предл.	0,012	−1359	2	30

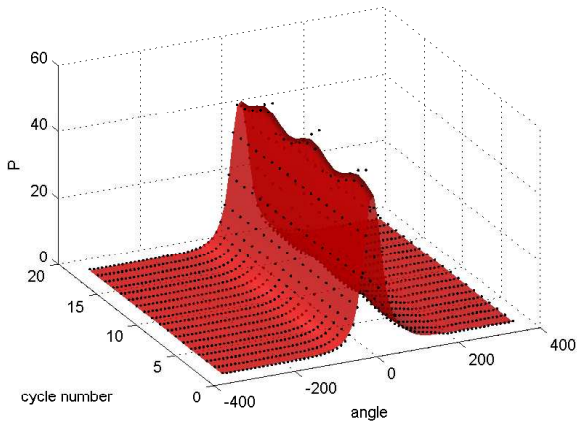


Рис. 4. Модель зависимости давления в камере двигателя внутреннего сгорания от угла поворота коленчатого вала и номера цикла. Точки соответствуют исходным данным.

внутреннего сгорания дизельного двигателя. Каждый временной ряд соответствует одному полному циклу работы двигателя, который состоит из рабочего и холостого тактов. Отчеты временного ряда равномерны и соответствуют углу вращения коленчатого вала. Нулевому углу соответствует верхняя мертвая точка вращения. Начало временного ряда соответствует углу в −360 градусов, конец — углу в +359,9 градусов. Всего один полный цикл насчитывает 7200 отсчетов. Лабораторный эксперимент включал измерения давления 122 полных циклов.

Регрессионная выборка

$$\left\{(\boldsymbol{\xi}^i, y^i)\right\}_{i=1}^m = \left\{(\langle n_i, \delta_i\rangle, P_i)\right\}_{i=1}^m$$

состоит из значений признаков n_i — номера измерения и δ_i — угла вращения коленчатого вала. Каждой паре значений n_i и δ_i, $i = 1, \ldots, m$ соответствует значение давления P_i в камере внутреннего сгорания. Задано множество порождающих функций

$$G = \left\{1/x,\ \sqrt{(|x|)},\ \exp\left(-\tfrac{(x-m_i)^2}{2s_i^2}\right),\ \sin(c_i x)\right\}.$$

Регрессионная выборка была случайным образом разбита на контрольную и обучающую, равные по мощности. Вычислялось значение оценки скользящего контроля CV для фиксированных 10 разбиений выборки, значение информационного критерия Акаике $\mathrm{AIC} = m\ln(S/m) + 2k$, десятичный логарифм числа обусловленности $\varkappa$ матрицы значений отобранных признаков и сложность модели k. Результаты экспериментов показаны в таблице 1. На рис. 2 и рис. 3 показаны соответственно график зависимости расстояния Макарова R_{ij} между моделями и график зависимости правдоподобия от номера итерации. На рис. 4 показана одна из полученных моделей.

Выводы

Для выбора наилучшей модели из индуктивно заданного множества использован двухуровневый Байесовский вывод. Предлагаемый алгоритм находит оптимальное решение в случае, показанном на рис. 1.

Предлагаемый алгоритм включает процедуру анализа мультиколлинеарности и позволяет получать хорошо обусловленные наборы порожденных признаков. Сравнение с LARS показало, что проверка мультиколлинеарности позволяет существенно уменьшить сложность получаемых моделей при незначительном ухудшении качества.

В результате работы алгоритма получена модель с удовлетворительной погрешностью аппроксимации, а матрица выбранных признаков имеет небольшую обусловленность, поэтому полученное решение устойчиво.

Литература

[1] *Vladislavleva E. J., Smits G. F., den Hertog D.* Order of nonlinearity as a complexity measure for models generated by symbolic regression via pareto genetic programming // IEEE Transactions on Evolutionary Computation. — 2009. — Vol. 13, No. 2. — Pp. 333–349.

[2] *Bishop C. M.* A new framework for machine learning // Computational Intelligence: Research Frontiers, IEEE World Congress on Computational Intelligence, WCCI 2008, Hong Kong. — Springer, 2008. — Pp. 1–24.

[3] *Draper N. R., Smith H.* Appied Regression Analysis. — John Wihley and Sons, 1998.

[4] *Marquardt D. W.* Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation // Technometrics. — 1996. — Vol. 12, No. 3. — Pp. 605–607.

[5] *Belsley D. A.* Conditioning Diagnostics: Collinearity and Weak Data in Reggression. — New York: John Wiley and Sons, 1991.

[6] *Efron B., Hastie T., Johnstone I., Tibshirani R.* Least angle regression // The Annals of Statistics. — 2004. — Vol. 32, No. 3. — Pp. 407–499.

[7] *Isenmann A. J.* Modern multivariate statistical techniques. — Springer, 2008. — 734 pp.

[8] *Макаров Л. И.* Метрические свойства функций расстояний между молекулярными графами // Журнал структурной химии. — 2007. — Vol. 48. — Pp. 223–229.

[9] *Tibshirani R.* Regression shrinkage and selection via the lasso // Journal of the Royal Statistical Society. — 1996. — Vol. 32, No. 1. — Pp. 267–288.