Алгоритм выбора нелинейных регрессионных моделей с анализом гиперпараметров*

Стрижов В. В., Сологуб Р. А.
strijov@ccas.ru, roman.sologub@yahoo.com
Москва, Вычислительный центр РАН, Московский физико-технический институт

Рассматривается задача порождения и выбора нелинейных регрессионных моделей. Модели индуктивно порождаются с помощью экспертно-заданного множества гладких функций. Для выбора моделей используется информация о распределении их параметров. Предлагается метод поиска моделей, комбинирующий подходы байесовского вывода и символьной регрессии.

Введение

Проблема выбора моделей является одной из наиболее актуальных и одновременно сложных при моделировании экономических, финансовых и социальных систем. Эта проблема состоит в отыскании оптимальной модели в некотором заданном классе моделей-претендентов. Этот класс задается в виде списка, либо в виде универсальной модели, из которой путем удаления элементов можно получить модели частного вида, либо с помощью правил порождения. В данной работе используется последний способ.

Критерий оптимальности модели задается исходя из гипотезы порождения данных — предположении о распределении случайной переменной при восстановлении регрессии и предположении о распределении параметров. Подход к модификации структуры моделей путем анализа параметров впервые предложен Ле Кюном и Хассиби в [1, 2]. Он состоит в исключении тех элементов моделей, мера выпуклости функции ошибки которых не превосходит заданный порог. Дальнейшее развитие методы анализа пространства параметров получили в работах Маккая [3, 4, 5, 6]. Им было предложено использовать гиперпараметры — параметры функций распределения данных и параметров для выбора моделей. В дальнейшем в работах [7, 8, 9] Бишоп предложил несколько способов оценки гиперпараметров: аппроксимацию Лапласа, ансамблевое обучение и оценку с помощью марковских цепей Монте-Карло.

Однако в рамках вышеприведенных подходов не рассматривались задачи порождения выбираемых моделей. Методы индуктивного порождения линейных регрессионных моделей описаны в работах Ивахненко [10, 11, 12]. Предложено порождать модели в виде линейных комбинаций мономов полинома Колмогорова–Габора. Методы порождения нелинейных регрессионных моделей развиты в работах Козы и Зелинки [13, 14]. Предложено порождать модели как произвольные суперпозиции заданного набора функций с помощью генетических оптимизированных алгоритмов. В работах Владиславлевой [15, 16] при выборе порождаемых моделей предлагается использовать Парето оптимальный фронт — множества моделей на плоскости, заданный функционал качества моделей и функций их сложности.

Ниже описан алгоритм, который является развитием алгоритма, опубликованного в [17, 18, 19]. Он выполняет следующие основные шаги. Задана модель начального приближения. Параметры этой модели настраиваются, вычисляются гиперпараметры, описывающие информативность элементов модели. Согласно гиперпараметрам, элементы модели модифицируются таким образом, чтобы с наибольшей вероятностью обеспечить попадание модели в Парето-оптимальный фронт.

Задача многомерной нелинейной регрессии

Задана регрессионная выборка — множество пар $D = \{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$, в котором $\boldsymbol{x} \in \mathbb{R}^P$ — свободная переменная, и $y \in \mathbb{R}^1$ — зависимая переменная.

Задано конечное множество порождающих функций $G = \{g \,|\, g\colon \mathbb{R} \times \ldots \times \mathbb{R} \to \mathbb{R}\}$. Функция $g = g(\boldsymbol{b}, \cdot, \ldots, \cdot)$ — гладкая параметрическая. Первый аргумент функции — вектор параметров, последующие аргументы — функции свободных переменных, принимающие значения в $\mathbb{R}^1$. Множество G индуктивно определяет набор допустимых суперпозиций $F = \{f_i\},\ i = 1, \ldots, M$. На эти суперпозиции накладывается ограничение сложности: каждая суперпозиция f_i состоит не более чем из R функций $g \in G$.

Суперпозиция f_i определяет параметрическую регрессионную модель $f_i = f_i(\boldsymbol{w}, \boldsymbol{x})$. Она зависит от независимых переменных $\boldsymbol{x}$ и вектора параметров $\boldsymbol{w}$. Вектор $\boldsymbol{w} \in \mathbb{R}^{W_i}$ состоит из присоединенных векторов — параметров функций $g_1, \ldots, g_{r_i}$, входящих в эту суперпозицию в лексикографическом порядке, то есть $\boldsymbol{w} = \boldsymbol{b}_1 \vdots \boldsymbol{b}_2 \vdots \ldots \vdots \boldsymbol{b}_{r_i}$, где $\vdots$ — знак присоединения векторов. Требуется отыскать в множестве F модель f_i, максимизирующую заданную целевую функцию $p(\boldsymbol{w} \,|\, D, A, \beta, f_i)$. Функция включает гиперпараметры A, β. Число параметров модели не должно превышать заданное число W^*. Число порождающих функций, из которых она состоит не должно превышать задан-

*Работа выполнена при финансовой поддержке РФФИ, проекты №07-07-00181, 08-01-12022.

ное число r^*. Модель, удовлетворяющую вышеперечисленным требованиям, будем называть моделью оптимальной структуры.

Распределение параметров моделей

Воспользуемся двухуровневым Байесовским выводом для оценки степени предпочтения порождаемых регрессионной моделей. Рассмотрим конечное множество моделей $f_1, \ldots, f_M$, приближающих данные D, обозначим априорную вероятность i-й модели $P(f_i)$. При появлении данных апостериорная вероятность модели $P(f_i \,|\, D)$ равна

$$P(f_i \,|\, D) = \frac{p(D \,|\, f_i)P(f_i)}{\sum_{j=1}^{M} p(D \,|\, f_j)P(f_j)}, \qquad (1)$$

где $p(D \,|\, f_i)$ — функция правдоподобия моделей, определяющая, насколько хорошо модель f_i описывает данные D. Знаменатель дроби обеспечивает выполнение условия $\sum_{i=1}^{M} P(f_i \,|\, D) = 1$.

Сравним две модели с помощью апостериорных вероятностей

$$\frac{P(f_i \,|\, D)}{P(f_j \,|\, D)} = \frac{p(D \,|\, f_i)P(f_i)}{p(D \,|\, f_j)P(f_j)}. \qquad (2)$$

Левая часть выражения называется отношением правдоподобия моделей. Отношение $P(f_i)/P(f_j)$ называется отношением апостериорных предпочтений моделей. Полагая априорные вероятности моделей одинаковыми, используем функции правдоподобия для выбора моделей.

Так как рассматриваемые модели f зависят от настраиваемых параметров, представим правдоподобие моделей в виде интеграла по пространству параметров

$$p(D \,|\, f) = \int p(D \,|\, \boldsymbol{w}, f)p(\boldsymbol{w} \,|\, f)d\boldsymbol{w}. \qquad (3)$$

Априорная плотность распределения параметров $\boldsymbol{w}$ модели f на выборке D равна

$$p(\boldsymbol{w} \,|\, D, f) = \frac{p(D \,|\, \boldsymbol{w}, f)p(\boldsymbol{w} \,|\, f)}{p(D \,|\, f)}, \qquad (4)$$

где $p(\boldsymbol{w} \,|\, f)$ — априорно заданная плотность вероятности параметров, и $p(D \,|\, \boldsymbol{w}, f)$ — функция правдоподобия параметров. Выражения (1) и (4) называются формулами Байесовского вывода первого и второго уровня.

Рассмотрим следующую гипотезу порождения данных при восстановлении регрессии

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

Пусть случайная величина ν имеет нормальное распределение $\mathcal{N}(0, \sigma^2)$ с нулевым матожиданием и дисперсией σ^2, которая не зависит от свободной переменной. Для фиксированной модели f плотность вероятности появления данных

$$p(y \,|\, \boldsymbol{x}, \boldsymbol{w}, \beta, f) \equiv p(D \,|\, \boldsymbol{w}, \beta, f) = \frac{\exp(-\beta E_D)}{Z_D(\beta)}, \quad (5)$$

где $\beta = \sigma^{-2}$, а коэффициент Z_D задан выражением, нормирующим функцию плотности в соответствии с гауссовым распределением

$$Z_D(\beta) = \left(\frac{2\pi}{\beta}\right)^{\frac{N}{2}}. \qquad (6)$$

Функция регрессионных невязок, согласно гипотезе порождения данных, равна

$$E_D = \frac{1}{2}\sum_{n=1}^{N} \left(f(x_n) - y_n\right)^2. \qquad (7)$$

Рассмотрим вектор параметров модели как многомерную случайную величину $\boldsymbol{w}$. Пусть плотность распределения параметров имеет вид многомерного нормального распределения $\mathcal{N}(\boldsymbol{0}, A)$ с матрицей ковариации A,

$$p(\boldsymbol{w} \,|\, A, f) = \frac{\exp(-E_{\boldsymbol{w}})}{Z_{\boldsymbol{w}}(A)}, \qquad (8)$$

где A — ковариационная матрица случайной величины $\boldsymbol{w}$. Нормирующая константа $Z_{\boldsymbol{w}}(A)$ равна

$$Z_{\boldsymbol{w}}(A) = (2\pi)^{\frac{W}{2}} |A|^{\frac{1}{2}}, \qquad (9)$$

где W — число параметров модели f. Функция-штраф за большое значение параметров модели при нормальном распределении равна

$$E_{\boldsymbol{w}} = \frac{1}{2}\boldsymbol{w}^{\intercal} A \boldsymbol{w}. \qquad (10)$$

При заданной модели f и заданных значениях A и β выражение (4) принимает вид

$$p(\boldsymbol{w} \,|\, D, A, \beta, f) = \frac{p(D \,|\, \boldsymbol{w}, \beta, f)p(\boldsymbol{w} \,|\, A, f)}{p(D \,|\, A, \beta, f)}. \qquad (11)$$

Записывая функцию ошибки

$$S(\boldsymbol{w}) = \frac{1}{2}\boldsymbol{w}^{\intercal} A \boldsymbol{w} + \beta E_D, \qquad (12)$$

получаем вместо (11) выражение

$$p(\boldsymbol{w} \,|\, D, A, \beta, f) \propto \frac{\exp(-S(\boldsymbol{w}))}{Z_S},$$

где Z_S — нормирующий множитель. Символ f далее будет опущен для удобства обозначений.

Вычисление гиперпараметров

Предлагается итеративно найти параметры и гиперпараметры модели по отдельности. На каждой итерации сначала при фиксированных гиперпараметрах отыскиваются параметры путем оптимизации функционала (12). Используется алгоритм Левенберга–Марквардта. Затем по формулам, предложенным ниже, вычисляются гиперпараметры.

Предположим, что после очередного шага итерации нам известен локальный максимум (12) и он находится в точке $\boldsymbol{w}_0$. Для нахождения гиперпараметров приблизим (11) методом Лапласа. Для этого построим ряд Тейлора второго порядка логарифма числителя (11) в окрестности $\boldsymbol{w}_0$

$$-S(\boldsymbol{w}) \approx -S(\boldsymbol{w}_0) - \frac{1}{2}\Delta\boldsymbol{w}^{\intercal} H \Delta\boldsymbol{w}, \qquad (13)$$

где $\Delta\boldsymbol{w} = \boldsymbol{w} - \boldsymbol{w}_0$. В выражении (13) нет слагаемого первого порядка, так как предполагается, что $\boldsymbol{w}_0$ доставляет локальный минимум функции ошибки

$$\left.\frac{\partial S(\boldsymbol{w})}{\partial \boldsymbol{w}}\right|_{\boldsymbol{w}=\boldsymbol{w}_0} = \mathbf{0}.$$

Матрица H — матрица Гессе функции ошибок

$$H = -\nabla\nabla S(\boldsymbol{w})|_{\boldsymbol{w}=\boldsymbol{w}_0}. \qquad (14)$$

Применяя экспоненту к обеим частям выражения (13) получаем требуемое приближение числителя (11)

$$\exp(-S(\boldsymbol{w})) \approx \exp(-S(\boldsymbol{w}_0)) \exp\left(-\tfrac{1}{2}\Delta\boldsymbol{w}^{\intercal} H \Delta\boldsymbol{w}\right). \qquad (15)$$

Учитывая то, что интеграл выражения (11) должен равняться единице, получаем нормирующий множитель

$$Z_S = \frac{\exp(-S(\boldsymbol{w}_0))(2\pi)^{\frac{W}{2}}}{|H|^{\frac{1}{2}}}. \qquad (16)$$

Знаменатель (11) является числителем (1) и определяет выбор наиболее правдоподобной модели. Для нахождения гиперпараметров максимизируем функцию $p(D \mid A, \beta)$ относительно A и β. Запишем ее в виде

$$p(D \mid A, \beta) = \int p(D \mid \boldsymbol{w}, A, \beta) p(\boldsymbol{w} \mid A) d\boldsymbol{w}. \qquad (17)$$

Используя выражения (5) и (8) перепишем (17) в виде

$$p(D \mid \beta, A) = \frac{Z_S}{Z_{\boldsymbol{w}}(A) Z_D(\beta)}.$$

Из (6), (9)и (16), логарифмируя (17), получим

$$\ln p(D \mid A, \beta) = -\frac{1}{2}\ln|A| - \frac{N}{2}\ln 2\pi + \frac{N}{2}\ln\beta - \\ - \beta E'_D - E'_{\boldsymbol{w}} - \frac{1}{2}\ln|H|. \qquad (18)$$

Найдем максимум выражения (18) относительно гиперпараметров, приравняв его производную поочередно по A и β к нулю. Для упрощения вычислений представим $A = \operatorname{diag}(\alpha) I_W$.

$$\frac{d\ln p(D \mid A, \beta)}{d\alpha} = -E'_{\boldsymbol{w}} + \frac{\boldsymbol{w}}{2\alpha} + \frac{d}{d\alpha}\ln\det(H).$$

Производная последнего слагаемого равна

$$\frac{d}{d\alpha}\ln|H| = \sum_{j=1}^{W} \frac{1}{\lambda_j + \alpha},$$

где λ_j — собственные значения матрицы H. Приравнивая последнее выражение к нулю и преобразовывая, получаем выражение для α

$$2\alpha E'_{\boldsymbol{w}} = W - \gamma, \quad \text{где} \quad \gamma = \sum_{j=1}^{W} \frac{\alpha}{\lambda_j + \alpha}. \qquad (19)$$

Аналогично получим β

$$2\beta E'_D = N - \sum_{j=1}^{W} \frac{\lambda_j}{\lambda_j + \alpha} = N - \gamma. \qquad (20)$$

Гиперпараметры α и β_i вычисляется итеративно следующим образом

$$\beta^{\text{new}} = \frac{N - \gamma}{E'_D}, \quad \alpha^{\text{new}} = \frac{W - \gamma}{E'_{\boldsymbol{w}}}.$$

Значения функционалов ошибок $E'_{\boldsymbol{w}}$ и E'_D оптимизируются после каждого вычисления новых значений гиперпараметров.

При выборе моделей выполняется следующая процедура. Экспертно задается модель-претендент. Каждому элементу модели ставится в соответствие свой гиперпараметр α. Параметры и гиперпараметры модели последовательно настраиваются. Элемент модели, имеющий наименьшее значение гиперпараметра, исключается. Модель пополняется новым элементом из множества G согласно заданному правилу. Так как на каждом шаге такой модификации модели функционал качества не ухудшается, процедура выполняется до сходимости функционала качества (13).

Вычислительный эксперимент

Проиллюстрируем итеративное изменение параметров и гиперпараметров с помощью модели $y = f_0(\boldsymbol{w}, \boldsymbol{x}) = w_1 + w_2 \sin x_1 + w_3 \sin x_2$. Свободные переменные данной модели имеют значения $x_1, x_2 \in \{0, 0.1, \ldots, 1\}$. Зависимые переменные, полученные как $y = f_0(\boldsymbol{w}, \boldsymbol{x}) + \nu_0$, где $\nu_0 \sim \mathcal{N}(0, \pi/2)$.

На рис. 1 показаны итеративные изменения параметров w_1, w_2, w_3, латентной переменной γ и гиперпараметра β. По оси абсцисс отложен номер

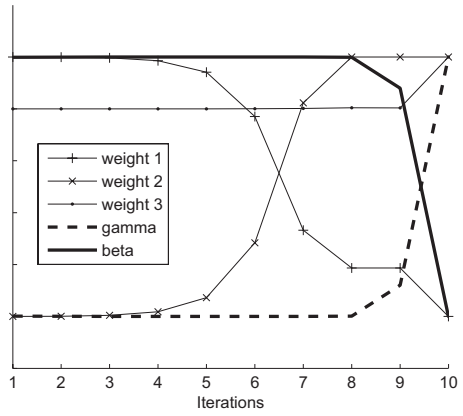

Рис. 1.

итерации. По оси ординат — нормированные значения переменных.

В результате эксперимента была получена сходящаяся последовательность, определяющая оптимальные параметры и гиперпараметры для данной модели. Итеративный алгоритм останавливается, когда значение функции ошибки текущей модели имеет значение меньше заданного, или же невозможна генерация новых моделей-претендентов.

Заключение

Данная работа описывает алгоритм выбора нелинейных регрессионных моделей из индуктивно-порождаемого множества. Для выбора моделей используются настраиваемые гиперпараметры. Каждый гиперпараметр ставится в соответствие элементы суперпозиции функций, задающих модель. По его значению определяется важность элемента суперпозиции и принимается решение о необходимости его модификации.

Литература

[1] *LeCun Y., Denker J., Solla S., Howard R. E., Jackel L. D.* Optimal brain damage // Advances in Neural Information Processing Systems II / edited by D. S. Touretzky. — San Mateo, CA: Morgan Kauffman, 1990. — P. 598–605.

[2] *Hassibi B., Stork D. G.* Second order derivatives for network pruning: Optimal brain surgeon // Advances in Neural Information Processing Systems / edited by S. J. Hanson, J. D. Cowan, C. L. Giles. — Vol. 5. — Morgan Kaufmann, San Mateo, CA, 1993. — P. 164–171.

[3] *MacKay D.* Information Theory, Inference, and Learning Algorithms. — Cambridge University Press, 2003. — 628 p.

[4] *Cavendish D. M., Mackay D. J.,C., Laboratory C.* Comparison of approximate methods for handling hyperparameters // Neural Computation. — 2003. — Vol. 11. — P. 1035–1068.

[5] *MacKay D. J.* Choice of basis for laplace approximation: Tech. rep.: Machine Learning, 1998.

[6] *Cawley G., Talbot N., Guyon I., Saffari A.* Preventing over-fitting during model selection using bayesian regularisation // Journal of Machine Learning Research. — 2007. — Vol. 8.

[7] *Bishop C.* Pattern Recognition And Machine Learning. — Springer, 2006.

[8] *Bishop C. M., Tipping M. E.* Bayesian regression and classification.

[9] *Barber D., Bishop C. M.* Ensemble learning in bayesian neural networks // Neural Networks and Machine Learning. — Springer, 1998. — P. 215–237.

[10] *Malada H. R., Ivakhnenko A. G.* Inductive Learning Algorithms for Complex Systems Modeling. — CRC Press, 1994. — 368 p.

[11] *Ивахненко А. Г., Юрачковский Ю. П.* Моделирование сложных систем по экспериментальным данным. — М.: Радио и связь, 1987. — 120 с.

[12] *Mueler J. A., Lemke F.* Sel-organising Data Mining: An Intelligent Approach To Extract Knowledge From Data. — Berlin: Dresden, 1999. — 225 p.

[13] *Koza J. R.* Genetic Programming IV: Routine Human-Competitive Machine Intelligence. — Springer, 2005.

[14] *Zelinka I., Nolle L., Oplatkova Z.* Analytic programming — symbolic regression by means of arbitrfary evolutionary algorithms // Journal of Simulation. — 2004. — Vol. 6(9). — P. 44–56.

[15] *Vladislavleva E.* Model-based Problem Solving through Symbolic Regression via Pareto Genetic Programming: PhD thesis. — Tilburg University, Tilburg, the Netherlands, 2008. — 288 p.

[16] *Vladislavleva E., Smith G., Hertog D.* Order of nonlinearity as a complexity measure for models generated by symbolic regression via pareto genetic programming // EEE Transactions on Evolutionary Computation. — 2009. — Vol. 13(2). — P. 333–349.

[17] *Стрижов В. В.* Поиск параметрической регрессионной модели в индуктивно заданном множестве // Журнал вычислительных технологий. — 2007. — № 1. — С. 93–102.

[18] *Стрижов В. В.* Методы индуктивного порождения регрессионных моделей. — М.: ВЦ РАН, 2008. — 54 с.

[19] *Стрижов В. В., Сологуб Р. А.* Индуктивное порождение регрессионных моделей предполагаемой волатильности для опционных торгов // Журнал вычислительных технологий. — 2009. — Т. 3.