УДК 519.2
В.В. Стрижов
Вычислительный центр им. А. А. Дородницына РАН
г. Москва
strijov@ccas.ru

ПОИСК РЕГРЕССИОННЫХ МОДЕЛЕЙ В ИНДУКТИВНО ЗАДАННОМ МНОЖЕСТВЕ *

АННОТАЦИЯ

Задана выборка значений нескольких свободных и одной независимой переменной и задан набор порождающих гладких функций, индуктивно определяющих множество регрессионных моделей. Описан алгоритм выбора оптимальной регрессионной модели, использующий гиперпараметры для анализа элементов модели.

ВВЕДЕНИЕ

Для построения регрессионной модели, которая оптимально описывает измеряемые данные, предлагается рассмотреть суперпозиции набора порождающих функций. Такой способ получения моделей описывается в работах, посвященных методу группового учета аргументов [1] и в работах по генерации структурных моделей [2]. Для порождения моделей используются генетические и эволюционные алгоритмы, обзор которых представлен в [3]. Наиболее близкой работой по данной тематике является [4]. Поиск моделей выполняется по итерационной схеме «порождение-выбор» в соответствии с правилами порождения моделей и критерием выбора моделей, определенных ниже.

Предлагаемый алгоритм, используя множество порождающих функций и множество моделей начального приближения, итеративно порождает набор моделей, среди которых, по заданному критерию качества, выбираются лучшие. При выборе моделей, элементам моделей ставятся в соответствие гиперпараметры, которые определяют информативность этих элементов. Метод отсечения неинформативных элементов, предложенный в работе [5], используется для модификации порождаемых моделей.

* Работа поддержана грантом РФФИ 04-01-00401.

Наименее информативные элементы изымаются из модели без существенного ухудшения ее качества. При этом количество параметров модели уменьшается. Выбор моделей выполняется в соответствии с критерием, который определяется посредством назначения модели данных.

МОДЕЛЬ ДАННЫХ

Рассмотрим регрессию $y = f(\mathbf{w}, \mathbf{x}) + \nu$ с аддитивным Гауссовским шумом с дисперсией σ_ν и с нулевым матожиданием. Такая модель данных задает плотность вероятности появления данных $p(y \mid \mathbf{x}, \mathbf{w}, \beta, f) = p(D \mid \mathbf{w}, \beta, f) = \exp(-\beta E_D(D \mid \mathbf{w}, f)) Z_D^{-1}(\beta)$, где $\beta = \sigma_\nu^{-2}$ и $Z_D(\beta)$ – нормирующий множитель. Взвешенная функция ошибки в пространстве данных равна $\beta E_D = \frac{\beta}{2} \sum_{n=1}^{N} (f(\mathbf{x}_n) - y_n)^2.$

Введем регуляризующий параметр α, который отвечает за то, насколько хорошо модель должна соответствовать зашумленным данным. Функция плотности вероятности параметров $\mathbf{w}$ с заданным α имеет вид $p(\mathbf{w} \mid \alpha, f) = \exp(-\alpha E_W(\mathbf{w} \mid f)) Z_W^{-1}(\alpha),$ где $\alpha = \sigma_w^{-2}$ соответствует дисперсии распределения весов, $Z_W(\alpha)$ – нормирующий множитель. Ошибка $\alpha E_W = \frac{\alpha}{2} \| \mathbf{w} \|^2$ регуляризирует параметры, начисляя штраф за их чрезмерно большие значения. Записывая функцию ошибки в виде $S(\mathbf{w} \mid f) = \alpha E_W + \beta E_D$ получаем $p(\mathbf{w} \mid D, \alpha, \beta, f) = \exp(-S(\mathbf{w} \mid f)) Z_S^{-1}(\alpha, \beta),$ где $Z_S(\alpha, \beta)$ – нормирующий множитель. Переменные α и β называются гиперпараметрами.

ПОСТАНОВКА ЗАДАЧИ

Задана выборка – множество $\{\mathbf{x}_1, ..., \mathbf{x}_N \mid \mathbf{x} \in \mathbf{R}^M\}$ значений свободных переменных и множество $\{y_1, ..., y_N \mid y \in \mathbf{R}\}$ соответствующих им значений зависимой переменной. Обозначим оба множества как множество исходных данных D. Выборка отвечает принятой модели распределения данных и рассматривается как результат измерений при проведении некоторого эксперимента, модель которого неизвестна.

Задано множество $G = \{g_1,...,g_K \mid g : \mathbf{R}\times...\times\mathbf{R} \to \mathbf{R}\}$ гладких параметрических функций, в котором $g = g(\mathbf{b},\cdot,...,\cdot)$, где первый аргумент – вектор-строка параметров функции, последующие – переменные из $\mathbf{R}$, рассматриваемые как элементы вектора свободных переменных. Функции этого множества называются порождающими.

Множество G задает множество $\Phi^* = \{f_i(\mathbf{w},\mathbf{x})\}$ путем обобщенного индуктивного определения. Регрессионная модель f_i – произвольная суперпозиция порождающих функций. Вектор $\mathbf{w} \in \mathbf{R}^W$ есть присоединенные векторы параметров функций, входящих в суперпозицию. В дальнейшем будет рассматриваться множество Φ, элементы которого состоят не более чем из r порождающих функций.

Требуется найти такую модель $f_i \in G$ и такие ее параметры $\mathbf{w}$, которые доставляли бы максимум целевой функции $p(\mathbf{w} \mid D,\alpha,\beta,f_i)$.

ОПИСАНИЕ АЛГОРИТМА

Перед работой алгоритма заданы множество измеряемых данных D, множество порождающих функций G и начальный набор конкурирующих моделей $F_0 = \{f_1,...,f_L \mid f \in \Phi\}$. Задано значение r, ограничивающее число элементов суперпозиции. На каждой итерации алгоритма, по правилам, описанным ниже, порождается набор F_s, состоящий из $2L$ моделей. Порожденные модели модифицируются. Цель модификации моделей состоит в отсечении наименее информативных элементов. Так как гиперпараметр α – обратная величина дисперсии параметров, то по малому значению этого гиперпараметра возможно сделать вывод о том, что большие значения параметров $\mathbf{w}$ допустимы. Для оценки информативности, каждому элементу модели ставится в соответствие гиперпараметр. Наименее информативным считается элемент модели с наибольшим значением гиперпараметра. Такой элемент изымается из модели. Среди порожденных моделей, соответствии с принятым критерием оптимальности, выбираются L наилучших, которые используются в следующей итерации. Итерации повторяется заданное количество раз.

Для нахождения значений параметров и гиперпараметров используется приближение целевой функции методом Лапласа: $S(\mathbf{w}) \approx S(\mathbf{w}^{MP}) + 2^{-1}(\mathbf{w}-\mathbf{w}^{MP})^T A(\mathbf{w}-\mathbf{w}^{MP})$. Параметры $\mathbf{w}^{MP}$ доставляют максимум правдоподобия целевой функции. Матрица

Гессе функции ошибок определена как $A = \nabla^2 S(\mathbf{w}^{MP}) = \beta \nabla^2 E_D(\mathbf{w}^{MP}) + \alpha I.$ Гиперпараметры α, β находятся из выражений $2\alpha E_W^{MP} = W - \gamma \quad 2\beta E_D^{MP} = N - \gamma,$ в которых $\gamma = \sum_{i=1}^{W} \frac{\alpha}{\lambda_i + \alpha}$, λ_i – собственные значения матрицы A.

Порождение моделей выполняется следующим образом. Для каждой модели f_i из набора F_s с помощью процедуры нелинейной оптимизации [6] минимизируется целевая функция $S_i(\mathbf{w})$ и отыскиваются параметры $\mathbf{w}^{MP}$. Модели ставится с соответствие гиперпараметр β_i, характеризующий плотность распределения данных, а каждой функции из порождающей ее суперпозиции $\{g_{ij}(\mathbf{b})\}_{j=1}^{r_i}$ ставится в соответствие гиперпараметр α_{ij}, характеризующий плотность распределения вектора параметров. Находятся значения гиперпараметров: $\beta_i = (N - \gamma) E_D^{-1}(f_i(\mathbf{w}^{MP}))$ и $\alpha_{ij} = \gamma E_W^{-1}(g_{ij}(\mathbf{b}^{MP}))$.

После отыскания параметров и гиперпараметров моделей выполняются операции порождения моделей $f_1', \ldots, f_L'$ из F_s. Для этого в каждой f_i выбирается функция g_{ij} с наибольшим значением α_{ij}. Выбирается модель f_ξ из $F_s / \{f_i\}$ и ее функция $g_{\xi\zeta}$. Модель f_i' порождается путем замещения функции g_{ij} с ее аргументами на функцию $g_{\xi\zeta}$ с ее аргументами.

После порождения, каждая модель f_i' подвергается модификации. Из множества G по заданному закону распределения выбирается функция g_k и замещает функцию g_{ij} модели f_i'. При этом начальное значение гиперпараметра α_{ij} этой функции определяется как $\max_j \alpha_{ij}$.

Из F_s выбираются L функций, имеющих максимальное значение целевой функции, итерация алгоритма повторяется.

РЕЗУЛЬТАТЫ

Описанный алгоритм был использован для построения модели изменения давления в камере внутреннего сгорания дизельного двигателя. Свободная переменная – угол поворота коленчатого вала, зависимая переменная – давление, измеряемое датчиком, установленным в камере. Измерения, выполненные в течение одного цикла работы двигателя, содержали 4000 отсчетов. Всего было рассмотрено 118 циклов. Получена модель с шестнадцатью параметрами, состоящая из четырех порождающих функций, удовлетворительная с точки зрения экспертов.

ЛИТЕРАТУРА

1. Malada, H.R., Ivakhnenko, A.G. Inductive Learning Algorithms for Complex Systems Modeling. CRC Press, 1994.

2. Marenbach, P., Betterhausen, K. and Freyerm, S. Signal Path Oriented Approach for Generation of Dynamic Process Models // Genetic Programming 1996: Proceedings of the First Annual Conference, MIT Press, 1996. P. 327–332.

3. Yao X. A review of evolutionary artificial neural networks // International Journal of Intelligent Systems, 8(4), 1993. P. 39–67.

4. Nikolaev, N. Iba, H. Accelerated Genetic Programming of Polynomials, Genetic Programmimg and Evolvable Machines. Kluwer Academic Publ., vol.2(3), 2002. P. 231–257.

5. LeCun, Y., Denker, J. S., and Solla, S. A., Optimal brain damage // Touretzky, D.S., ed., Advances in Neural Information Processing Systems 2. Morgan Kaufmann, San Mateo, CA, 1990. P. 598–605.

6. Coleman, T.F., Li, Y., An Interior, Trust Region Approach for Nonlinear Minimization Subject to Bounds. SIAM Journal on Optimization, vol. 6, 1996. P. 418–445.

V.V. Strijov

Stable integral indicators with the choice of objects features for a support set

There a sample set of unbound variables and one variable are given. There a set of non-generated functions, which define a set of regression models, is given. The paper describes an algorithm of optimal regression model choice. The algorithm uses hyperparameters to estimate model elements importance.

В.В. Стрижов

Пошук регресійних моделей у индуктивно заданій безлічі

Задано вибірку значень декількох вільних і однієї незалежної перемінної і заданий набір непороджуваних гладкий функцій, индуктивно визначальну безліч регресійних моделей. Описано алгоритм вибору оптимальної регресійної моделі, использующий гіперпараметри для аналізу елементів моделі.