Математические методы распознавания образов. (ММРО-12). Доклады XII Всероссийской конференции. РАН, ВЦ, РФФИ. Москва, 2005.

Поиск модели оптимальной сложности в задачах нелинейной регрессии

Стрижов В.В.
(Москва)

Широкое практическое применение методов нелинейной оптимизации в регрессионном анализе подготовило базу для создания алгоритмов синтеза регрессионных моделей. Рассматривается процедура поиска оптимальной регрессионной модели в классе моделей, определенном суперпозициями гладких функций из заданного множества. Для поиска используются алгоритмы генетической оптимизации. Параметры моделей оцениваются с помощью методов нелинейной оптимизации.

Поставим задачу нахождения регрессионной модели нескольких свободных переменных следующим образом. Пусть 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. Эту выборку мы будем рассматривать как результат измерений при проведении некоторого эксперимента, модель которого неизвестна.

Также задано множество $G = \{g \mid g : \mathbf{R} \times ... \times \mathbf{R} \to \mathbf{R}\}$ параметрических гладких функций $g = g(\mathbf{b},\cdot,...,\cdot)$. Первый аргумент функции g – вектор-строка параметров $\mathbf{b}$, последующие – переменные из множества действительных чисел, интерпретируемые как элементы вектора свободных переменных. Рассмотрим произвольную суперпозицию f функций из G, которая задает параметрическую регрессионную модель $f = f(\mathbf{w}, \mathbf{x})$, где вектор $\mathbf{w} \in \mathbf{R}^W$ состоит из присоединенных векторов-параметров функций g. Обозначим $\Phi = \{f_i\}$ – счетное множество всех суперпозиций функций из G.

Требуется найти такую модель f_i, которая доставляет максимум функционала $p(\mathbf{w} \mid D, \alpha, \beta, f_i)$.

Общий подход к выбору нелинейных моделей описан МакКаем в [1] и заключается в следующем. Рассмотрим регрессию $y = f(\mathbf{w}, \mathbf{x}) + \nu$ с аддитивным Гауссовским шумом с дисперсией σ_ν и с нулевым математическим ожиданием. Тогда вероятность появления данных

$$p(y \mid \mathbf{x}, \mathbf{w}, \beta, f) = p(D \mid \mathbf{w}, \beta, f) = \frac{\exp(-\beta E_D(D \mid \mathbf{w}, f))}{Z_D(\beta)},$$

где $\beta = \sigma_\nu^{-2}$, $Z_D(\beta)$ – нормирующий множитель. Взвешенный функционал ошибки в пространстве данных

$$\beta E_D = \frac{\beta}{2}\sum_{n=1}^{N}(f(\mathbf{x}_n) - y_n)^2.$$

Точность аппроксимации данных зависит, вообще говоря, от сложности модели и задача оценки наиболее правдоподобных весов в сложных моделях не является корректной в смысле Адамара. Введем регуляризующий параметр α, который отвечает за то, насколько хорошо модель должна соответствовать зашумленным данным.

Функция вероятности весов с заданным параметром α имеет вид

$$p(\mathbf{w} \mid \alpha, f) = \frac{\exp(-\alpha E_W(\mathbf{w} \mid f))}{Z_W(\alpha)},$$

где $\alpha = \sigma_w^{-2}$ соответствует дисперсии распределения весов. Ошибка $\alpha E_W = \frac{\alpha}{2} \| \mathbf{w} \|^2$ регуляризирует веса, начисляя штраф за их чрезмерно большие значения.

Записывая функцию ошибки в виде $S(w \mid f) = \alpha E_W + \beta E_D$ получаем

$$p(\mathbf{w} \mid D, \alpha, \beta, f) = \frac{\exp(-S(\mathbf{w} \mid f))}{Z_S(\alpha, \beta)}.$$

Для нахождения этого функционала используется приближение рядом Тейлора второго порядка: $S(\mathbf{w}) \approx S(\mathbf{w}^{MP}) + \frac{1}{2}(\mathbf{w} - \mathbf{w})^T A(\mathbf{w} - \mathbf{w})$. Матрица Гессе функции ошибок определена как $A = \nabla^2 S(\mathbf{w}^{MP}) = \beta \nabla^2 E_D(\mathbf{w}^{MP}) + \alpha I$.

Проблема выбора свободных переменных на каждом элементе суперпозиции является общей проблемой для задач распознавания образов. Для линейных моделей существует хорошо разработанная процедура анализа главных компонент, в которой наиболее важные переменные имеют большую корреляцию с первой главной компонентой. Ле Кюн [2] предложил подобный метод для нелинейных моделей. Он заключается в анализе матрицы Гессе и использует алгоритм сокращения весов. В нем отдельные гиперпараметры связаны с группами весов. В течение обучения гиперпараметры изменяются. По их значению возможно сделать вывод о важности данной свободной переменной.

Гиперпараметры α, β находятся из выражений $2\alpha E_W^{MP} = W - \gamma$ и $2\beta E_D^{MP} = N - \gamma$, где $\gamma = \sum_{i=1}^{W} \frac{\alpha}{\lambda_i + \alpha}$ и λ_i – собственные значения матрицы A.

Поиск оптимальной модели происходит на множестве моделей, порождаемых на каждой итерации генетического оптимизационного алгоритма. Перед работой алгоритма заданы множество данных D и

множество G гладких на области определения свободной переменой функций. Задан начальный набор конкурирующих моделей, $F_0 = \{f_1,...,f_L \mid f \in \Phi\}$, в котором каждая модель f_i есть суперпозиция функций $\{g_{ij}\}_{j=1}^{r_i}$. Модели f_i поставлен в соответствие гиперпараметр β_i, функции g_{ij} поставлен в соответствие гиперпараметр α_{ij}, имеющие значения начального приближения. Далее выполняется последовательность шагов, которые повторяются заданное количество раз.

Для каждой модели из набора F минимизируем штрафные функции $S_i(\mathbf{w})$ с помощью процедуры нелинейной оптимизации и получаем веса $\mathbf{w}^{\mathrm{MP}}$. Определим новые значения гиперпараметров α_{ij} и β_i. Изменение гиперпараметров может быть повторено итерационно после нахождения нового локального минимума $S_i(\mathbf{w})$. После отыскания гиперпараметров моделей выполняются стандартные операции кроссовера, мутации и селекции.

Заданы правила построения производных моделей $f_1',...,f_L'$. В f_i выбирается функция g_{ij} с наименьшим значением α_{ij}. Выбирается модель f_ξ из $F/\{f_i\}$ и ее функция $g_{\xi\zeta}$. Модель f_i' порождается из модели f_i путем замещения функции g_{ij} с ее аргументами на функцию $g_{\xi\zeta}$ с ее аргументами.

С заданной вероятностью каждая порожденная модель подвергается мутации. Из множества G случайным образом выбирается функция g' и замещает функцию g_j мутирующей модели. Гиперпараметр α_{ij} этой функции определяется как $\min_j \alpha_{ij}$. Вектор параметров этой функции $\mathbf{b}_{ij}$ назначается при задании G.

Из объединенного множества родительских и порожденных функций в соответствии с критерием S выбираются M наилучших, которые участвуют в дальнейшей оптимизации.

Предлагаемый метод протестирован на задаче по моделированию процесса горения в камере внутреннего сгорания дизельного двигателя.

Данная работа поддержана грантом РФФИ 04-01-00401-а.

Литература

1. MacKay D., Information, inference, learning algorithms. Cambridge University Press, 2003.
2. LeCun, et al., Optimal brain damage. //Touretzky, D.S., Advances in Neural Information Processing Systems 2, p. 598–605., San Mateo, CA, 1990.