Алгоритмы индуктивного порождения и критерии выбора оптимальной существенно нелинейной регрессионной модели

Г. И. Рудой

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем
Научный руководитель: В. В. Стрижов

Июнь 2014

- Разработка и анализ алгоритма порождения суперпозиций существенно нелинейных регрессионных моделей.
  - Доказательство существования данной суперпозиции.
  - Разработка практически реализуемого алгоритма.
- Формулирование и обоснование понятия устойчивости параметров моделей:
  - Критерий выбора моделей.
  - Исследование погрешности в измеряемых данных.

Постановка задачи выбора модели

Дана выборка

$$D = \{(\mathbf{x}_i, y_i) \mid i \in \{1, \dots, N\}, \mathbf{x}_i \in \mathbb{X} \subset \mathbb{R}^n, y_i \in \mathbb{Y} \subset \mathbb{R}\}.$$

Для множества всех суперпозиций

$$\mathcal{F} = \{f_r \mid f_r : (\boldsymbol{\omega}, \mathbf{x}) \mapsto y \in \mathbb{Y}, r \in \mathbb{N}\},$$

требуется найти индекс $\hat{r}$ такой, что функция $f_{\hat{r}}$ доставляет минимум функционалу качества Q:

$$\hat{r} = \arg\min_{r \in \mathbb{N}} Q(f_r \mid \hat{\boldsymbol{\omega}}_{\boldsymbol{r}}, D),$$

$$\hat{\boldsymbol{\omega}}_{\boldsymbol{r}} = \arg\min_{\boldsymbol{\omega} \in \Omega} S(\boldsymbol{\omega} \mid f_r, D).$$

Аппроксимация выборки некоторой формулой:

$$y = \sin x_1^2 + 2x_2.$$

Методы:

- Генетические алгоритмы [Koza1998].
- Аналитическое программирование [Zelinka2008, Webb2010].

Проблемы аналитического программирования

- Порождение рекурсивных суперпозиций:
$$f = y + f(x, y).$$
- Несоответствие арности функций числу и типам аргументов:
$$f = \sin(x, y, z).$$
- Несовпадение областей определения и значений:
$$f = \sqrt{-x^2}.$$
- Порождение слишком сложных суперпозиций.

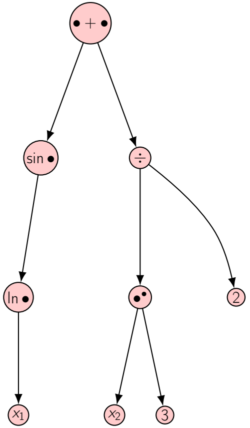
Постановка теоретической задачи

Пусть $G = \{g_1, \dots, g_l\}$ — множество данных порождающих функций; для каждой $g_i \in G$ заданы:

- функция (например, sin, cos, $\times$),
- арность функции и порядок следования аргументов,
- тип аргументов ($\mathrm{dom} g_i$) и тип значения ($\mathrm{cod} g_i$) функции,
- область определения $\mathcal{D} g_i \subset \mathrm{dom} g_i$ и область значений $\mathcal{E} g_i \subset \mathrm{cod} g_i$.

Требуется:

- построить алгоритм, за конечное число итераций порождающий любую конечную суперпозицию данных примитивных функций,
- оценить сложность полученного алгоритма,
- доказать полноту полученного алгоритма.



$$f = \sin(\ln x_1) + \frac{x_2^3}{2}.$$

$$G = G_b \cup G_u; X = \{x_1, \ldots, x_n\}.$$

1. Инициализация:
$$\mathcal{F}_0 = X,$$
$$\mathcal{I} = \{(x, 0) \mid x \in X\}.$$
2. Вспомогательные множества:
$$U_i = \{g_u \circ f \mid g_u \in G_u, f \in \mathcal{F}_i\},$$
$$B_i = \{g_b \circ (f, h) \mid g_b \in G_b, f, h \in \mathcal{F}_i\}.$$
3. $\mathcal{F}_{i+1} = \mathcal{F}_i \cup U_i \cup B_i$.
4. $\mathcal{I} = \mathcal{I} \cup (f, i+1)$, если f не присутствует в $\mathcal{I}$.

Множество всех возможных суперпозиций $\mathcal{F} = \cup_{i=0}^{\infty} \mathcal{F}_i$.

Теорема

Предложенный алгоритм породит любую конечную суперпозицию за конечное число шагов.

Теорема

Пусть в множестве примитивных функций G *содержится* l_p *функций арности* $p > 1$ *и ни одной функции арности* $p + k \mid k > 0$, *и имеется* $n > 1$ *независимых переменных. Тогда справедлива следующая оценка количества суперпозиций, порожденных предложенным алгоритмом после* k*-ой итерации:*

$$|\mathcal{F}_k| = \mathcal{O}(l_p^{\sum_{i=0}^{k-1} p^i} n^{p^k}).$$

- Суперпозиции порождаются случайным образом.
- Наименее удачные суперпозиции изменяются с сохранением структуры.
- Наиболее удачные суперпозиции комбинируются.

Критерий выбора моделей

$$Q_f = \frac{1}{1 + S_f} \left(\alpha + \frac{1 - \alpha}{1 + \exp(\frac{C_f}{\beta} - \tau)} \right).$$

- S_f — функционал ошибки.
- C_f — сложность модели.
- α — коэффициент влияния штрафа за сложность, $0 \ll \alpha < 1$,
- β — коэффициент строгости штрафа за сложность, $\beta > 0$,
- τ — коэффициент, характеризующий желаемую сложность модели.

Дано:

- Обучающая выборка D:

$$D = \{\mathbf{x}_i, y_i\} \mid i \in \{1, \dots, \ell\}.$$

- Семейство $\mathcal{F}$ параметрических функций $f = f(\mathbf{x}, \boldsymbol{\omega})$.
- Функционал качества S:

$$S = S(f(\cdot, \boldsymbol{\omega}), D) \mid f \in \mathcal{F}, \quad S \to \min_{\boldsymbol{\omega}}$$

Требуется:

- Исследовать зависимость $\hat{\boldsymbol{\omega}} = \arg\min_{\boldsymbol{\omega}} S$ от вариации D.
- Выбрать оптимальную модель согласно зависимости от вариации D.
- Проверить возможность экспертного применения f при данной вариации D.

- Случай линейной регрессии:

$$y_i = ax_i + b + \xi_i \mid i \in \{1, \dots, n\}, \xi_i \in \mathcal{N}(0, \sigma).$$

- Влияние пертурбаций на решения оптимизационных задач [Bonnans1998].
- Верхние границы ошибок в SVM [Vapnik2000].
- Сравнение стабильности и обобщающей способности при варьировании обучающей выборки [Bousquet2002].
- Вычислительная стабильность интерполяции [Higham2003].
- Стабильность алгоритмов кластеризации [Luxburg2009].
- Малые изменения входных данных в сетях глубокого обучения [Szegedy2014].

Предлагаемый алгоритм

1. Фиксируется параметрическая модель $f \in \mathcal{F}$:

$$f = f(\mathbf{x}, \omega) \in \mathcal{F}.$$

2. Начальный оптимальный вектор параметров:

$$\hat{\omega}_f(D) = \arg\min_{\omega_f} S(f, D).$$

Предлагаемый алгоритм

3. Варьируется выборка:

$$\acute{D}(\Sigma^{\mathbf{x}}, \boldsymbol{\sigma}^y) = \{\mathbf{x}_i + \boldsymbol{\xi}_i^{\mathbf{x}}, y_i + \xi_i^y \mid i \in 1, \dots, \ell; \\ \boldsymbol{\xi}_i^{\mathbf{x}} \sim \mathcal{N}(0; \boldsymbol{\sigma}_{i\cdot}^{\mathbf{x}}); \\ \xi_i^y \sim \mathcal{N}(0; \sigma_i^y)\},$$

где $\Sigma^{\mathbf{x}} = \|\boldsymbol{\sigma}_{ij}^{\mathbf{x}}\|$.

4. Оптимальный вектор параметров для варьированной выборки $\acute{D}$:

$$\hat{\boldsymbol{\omega}}_f(\acute{D}(\Sigma^{\mathbf{x}}, \boldsymbol{\sigma}_y)) = \arg\min_{\boldsymbol{\omega}_f} S(f(\cdot, \boldsymbol{\omega}_f), \acute{D}(\Sigma^{\mathbf{x}}, \boldsymbol{\sigma}_y)).$$

5. Разность с начальным оптимальным вектором $\hat{\boldsymbol{\omega}}_f$:

$$\Delta\hat{\boldsymbol{\omega}}_f(\acute{D}(\Sigma^{\mathbf{x}}, \boldsymbol{\sigma}_y)) = \hat{\boldsymbol{\omega}}_f(D) - \hat{\boldsymbol{\omega}}_f(\acute{D}(\Sigma^{\mathbf{x}}, \boldsymbol{\sigma}_y))$$

6. Шаги 3-5 повторяются N раз:

$$\acute{\mathcal{D}}_N(\Sigma^{\mathbf{x}}, \boldsymbol{\sigma}_y) = \{\acute{D}_1(\Sigma^{\mathbf{x}}, \boldsymbol{\sigma}_y), \dots, \acute{D}_N(\Sigma^{\mathbf{x}}, \boldsymbol{\sigma}_y)\}.$$

6. Вычисляется стандартное отклонение каждой компоненты вектора параметров:

$$\sigma_{\omega_i} = \mathrm{stddev}((\Delta \hat{\boldsymbol{\omega}}_f)_i).$$

7. Устойчивость i-го параметра относительно компоненты j описания:

$$T_f^N(i, j, \Sigma^{\mathbf{x}}, \boldsymbol{\sigma}_y) = \frac{\frac{\sigma_{\omega_i}}{\hat{\omega}_i}}{r(\{\frac{\boldsymbol{\sigma}^{\mathbf{x}}_{kj}}{\mathbf{x}_{kj}}\}_{k=1}^{\ell})}.$$

r выбирается экспертом, например:

- $r(a_1, \dots) = a_1$ — для равных относительных погрешностей;
- $r(a_1, \dots, a_\ell) = \frac{\sum_{i=1}^{\ell} a_i}{\ell}$ — средняя относительная погрешность.

$T > 1 \Rightarrow$ относительная погрешность параметра больше относительной погрешности в данных.

Теорема (Рудой)

Пусть стандартные отклонения j-ых компонент $\mathbf{x}$ одинаковы:

$$\forall i_1, i_2, j : \sigma^{\mathbf{x}}_{i_1 j} = \sigma^{\mathbf{x}}_{i_2 j}.$$

Пусть коэффициенты ω_i попарно не коррелируют:

$$\forall i_1, i_2 : Cov(\omega_{i_1}, \omega_{i_2}) = 0.$$

Тогда для достаточно малых $\sigma^{\mathbf{x}}_{ij}$:

$$\begin{aligned}\{\sigma_{\omega_i}(\boldsymbol{\sigma}_{\cdot 1}, \boldsymbol{\sigma}_{\cdot 2}, \dots, \boldsymbol{\sigma}_{\cdot |\mathbf{x}|})\}^2 = {} & \{\sigma_{\omega_i}(\boldsymbol{\sigma}_{\cdot 1}, 0, 0, \dots, 0)\}^2 + \\ & + \{\sigma_{\omega_i}(0, \boldsymbol{\sigma}_{\cdot 2}, 0, \dots, 0)\}^2 + \\ & + \dots + \\ & + \{\sigma_{\omega_i}(0, 0, \dots, 0, \boldsymbol{\sigma}_{\cdot |\mathbf{x}|}\}^2 + O(\sigma^{\mathbf{x}}_{ij}).\end{aligned}$$

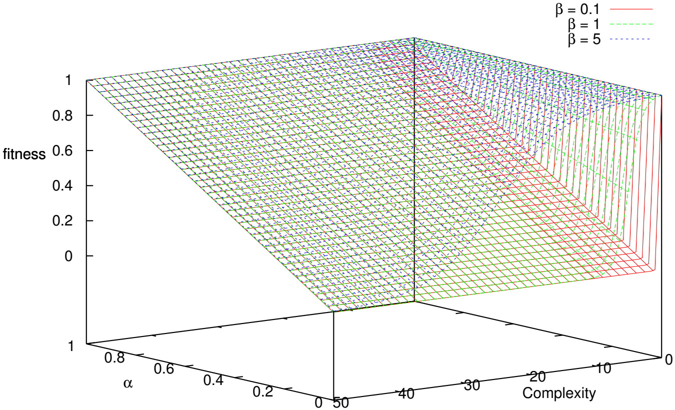
Дано:

- $D_j = (\lambda_i^j, n_i^j) \mid i \in \{1, \ldots, 17\}, j \in \{1, 2\}$.
- Экспертные предположения.

Требуется:

- $n_j = n_j(\lambda)$.
- Оценить адекватность $n_1(\lambda) - n_2(\lambda)$.

$$Q(f) = \frac{1}{1 + S(f)} \left(\alpha + \frac{1 - \alpha}{1 + \exp(\frac{C(f)}{\beta} - \tau)} \right).$$



Две модели:

- $n_1(\lambda) = 1.34 + \frac{3.54 \cdot 10^3}{\lambda^2} + \frac{2 \cdot 10^3}{\lambda^4}$.
- $n_2(\lambda) = 1.34 + \frac{11.6}{\lambda} + \frac{17.37}{\lambda^2} + \frac{0.0866}{\lambda^3} + \frac{2.95 \cdot 10^{-4}}{\lambda^4} + \frac{8.54 \cdot 10^{-7}}{\lambda^5}$.

τ	Суперпозиция	MSE	$C(f)$	$Q(f)$
10	n_1	$2.4 \cdot 10^{-8}$	13	0.095
30	n_2	$3.9 \cdot 10^{-9}$	31	0.031

Экспертное мнение: n_2 некорректна, нечетных степеней быть не может.

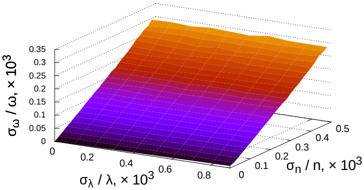
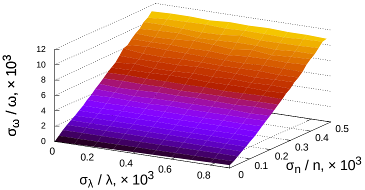
n_1	n_2
	

Таблица: Графики стандартного отклонения первого коэффициента для моделей n_1 и n_2.

Устойчивость второй модели в ≈ 40 раз хуже.

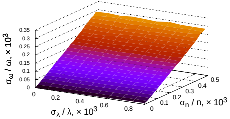
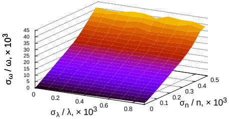
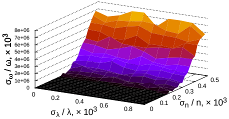
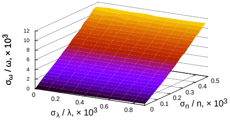
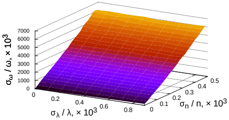
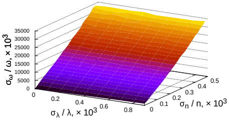
n	ω_1	ω_2	ω_3
1			
2			

Таблица: Графики стандартного отклонения первых трех коэффициентов для моделей n_1 и n_2.

Полимер	ω_1	ω_2	ω_3	MSE
1	1.34946	3558.95	1924.33	$2.2 \cdot 10^{-8}$
2	1.34047	3118.84	1578.59	$1.4 \cdot 10^{-8}$
Разность	$6.71 \cdot 10^{-3}$	$1.41 \cdot 10^{-1}$	$2.2 \cdot 10^{-1}$	

Таблица: Значения коэффициентов для модели n_1 и их относительная разность.

Коэфф.	$(2 \cdot 10^{-4}; 2 \cdot 10^{-5})$	$(6 \cdot 10^{-4}; 6 \cdot 10^{-5})$	$(9 \cdot 10^{-4}; 2 \cdot 10^{-4})$
1	$1.22 \cdot 10^{-5}$	$3.59 \cdot 10^{-5}$	$1.19 \cdot 10^{-4}$
2	$1.48 \cdot 10^{-3}$	$4.38 \cdot 10^{-3}$	$1.44 \cdot 10^{-2}$

Таблица: Значения стандартного отклонения для коэффициентов модели n_1 для первого полимера в зависимости от относительных дисперсий $(\frac{\sigma_\lambda}{\lambda}, \frac{\sigma_{\boldsymbol{n}}}{\boldsymbol{n}})$.

Модели разделимы: $\omega_1^1 - \omega_1^2 = 6.71 \cdot 10^{-3} >> \sigma_{\omega_1} = 1.19 \cdot 10^{-4}$.

Существенно переобученная модель:

$$L(x) = \prod_{i=0}^{\ell} y_i \prod_{j=0, j\neq i}^{\ell} \frac{x - x_j}{x_i - x_j},$$

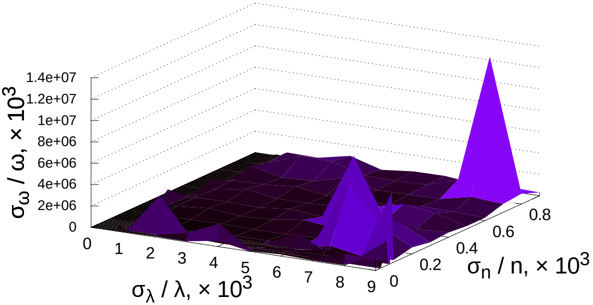


Рис.: Поверхность стандартного отклонения коэффициента ω_0.

Публикации и работы

- Г. И. Рудой. Анализ устойчивости существенно нелинейных регрессионных моделей к погрешностям в измеряемых данных. — «ЖВММФ» (направлено в журнал).
- Г. И. Рудой. О возможности применения методов Монте-Карло в анализе нелинейных регрессионных моделей. — «СибЖВМ» (направлено в журнал).
- Г. И. Рудой. Исследование устойчивости существенно нелинейных регрессионных моделей к погрешностям в обучающей выборке. — Труды 56-й научной конференции МФТИ. Раздел «Управление и прикладная математика», т. 1, с. 102-103. Москва, Долгопрудный, 2013.
- Г. И. Рудой. Устойчивость существенно нелинейных регрессионных моделей и метод её исследования. — Труды международной конференции студентов, аспирантов и молодых ученых «ЛОМОНОСОВ-2014», секция «Математическая статистика и ее приложения». Москва, 2014.

Результаты

- Предложено понятие устойчивости параметров модели.
- Обосновано использование понятия устойчивости параметров модели в качестве критерия выбора моделей.
- Продемонстрировано использование понятия устойчивости для анализа применимости экспертных моделей.
- Исследованы различные регрессионные модели и связь предложенного критерия с критериями ошибки и сложности модели.