Выбор предсказательной модели в режиме многозадачного обучения с применением символьных методов

Набиев Мухаммадшариф Фуркатович

Московский физико-технический институт
Кафедра интеллектуальных систем ФПМИ МФТИ

2025

Цель исследования

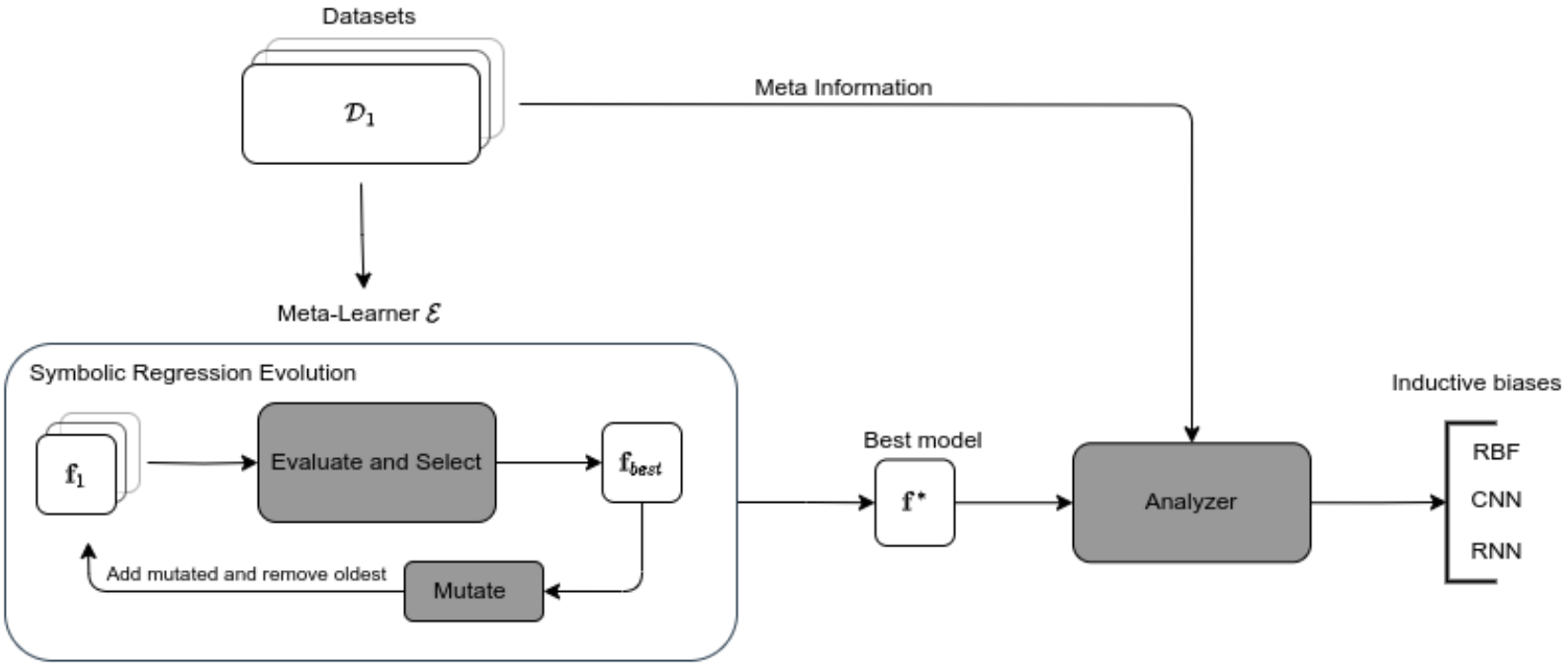
Проблема: Выбор оптимальной предсказательной модели сильно зависит от априорного знания человека о природе данных, т.е. от их индуктивного смещения. Определение индуктивного смещения автоматическим образом является открытой проблемой.

Примером индуктивного смещения могут служить сверточные нейронные сети для картинок.

Цель: Предложить метод автоматического определения индуктивного смещения.

Решение: Построение модели в режиме многозадачного обучения с применением символьной регрессии и дальнейшая ее интерпретация.

Архитектура решения



Мета-алгоритм $\mathcal{E}$ принимает на вход наборы данных и эволюционным путем строит модель. Далее лучший кандидат анализируется и делается вывод об индуктивном смещении.

Постановка задачи

Пусть $\mathfrak{T} = \{T_i\}_{i=1}^n$ – множество задач. Каждой задаче T_i соответствует набор данных $\mathfrak{D}_i = \{(\mathbf{x}_j, \mathbf{y}_j)\}_{j=1}^{N_i}$. Также обозначим $\mathfrak{D} = \{\mathfrak{D}_i\}_{i=1}^n$ и $\mathfrak{F}$ – множество всех моделей.

- Моделью $\mathbf{f}$ назовем отображение из $X \to Y$, структура Γ модели которой представима в виде дерева разбора символьного выражения.
- Параметрическим множеством декомпозируемых моделей $\mathfrak{F}$ назовем множество моделей с одинаковой структурой, которые представимы в виде $\mathbf{f} = \mathbf{g} \circ \mathbf{h}$.
- Назовем *индуктивным смещением* для декомпозируемой модели структуру первой части $\mathbf{h}$ модели $\mathbf{f}$.
- Имея функцию потерь задачи $\mathcal{L}_{\text{task}}$, мы хотим найти индуктивное смещение для декомпозируемой модели по заданному набору выборок:

$$\Gamma = \arg\min_{\mathbf{f}=\mathbf{g}\circ\mathbf{h}} \mathbb{E}_{\mathbf{X},\mathbf{Y}} \mathcal{L}_{\text{task}}(\mathbf{f}(\mathbf{X}), \mathbf{Y})$$

Данные для эксперимента

Эксперимент проводился на двух задачах:

1. **Бинарная классификация по окружности**: разделение точек, граница классов которых представляет собой окружность. Оптимальной моделью является *кривая второго порядка*.
2. **Авторегрессия**: предсказание временного ряда, сгенерированного и использованием авторегрессионной модели. Оптимальная модель находит *временные зависимости*.

Гипотеза: символьное представление модели будет содержать операции характерные для типа задачи.

Результаты эксперимента

Выборка	Кол-во выборок	Модели
Окружности	1	$g(x) = x + c_1,\ h(x,y) = (x + c_2)^2 + (y + c_3)^2$
	2	$g(x) = x + c_1,\ h(x,y) = (x + c_2)^2 + (y + c_3)^2$
	5	$g(x) = x + c_1,\ h(x,y) = \frac{xy(x+c_2)(x-y)}{c_3-y}$
	10	$g(x) = \frac{c_1}{x} + c_2,\ h(x,y) = \frac{y}{xc_3} - (x + c_4)(y + c_5)$
Авторегрессия	1	cx_t
	2	cx_t
	5	cx_t
	10	$csin^2(x_t)x_t$

Модели для разных размеров выборок.

Вывод

Заключение:

- Предложен метод автоматического определения индуктивного смещения в моделях.
- Проведены предварительные эксперименты на ряде синтетических выборок различного типа.
- Эксперименты показали применимость метода.

Дальнейшие исследования:

- **Бинарная классификация цифр (MNIST)**: выявление формулы с ограниченным числом параметров, в которой можно выделить компоненты, аналогичные свёрточными операциями
- Интеграция Informational Bottleneck для нахождения оптимальных (в смысле информации) моделей.

Источники

[1] A. I. Bazarova, A. V. Grabovoy, and V. V. Strijov. Analysis of the properties of probabilistic models in expert-augmented learning problems. *Autom. Remote Control*, 83(10):1527–1537, October 2022.

[2] M. Cranmer. Interpretable machine learning for science with pysr and symbolicregression.jl. *ArXiv*, abs/2305.01582, 2023.

[3] Miles Cranmer, Alvaro Sanchez-Gonzalez, Peter Battaglia, Rui Xu, Kyle Cranmer, David Spergel, and Shirley Ho. Discovering symbolic models from deep learning with inductive biases. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[4] Esteban Real, Chen Liang, David So, and Quoc Le. Automl-zero: Evolving machine learning algorithms from scratch. In *International conference on machine learning*, pages 8007–8019. PMLR, 2020.