Выбор предсказательной модели в режиме многозадачного обучения с применением символьных методов

Набиев Мухаммадшариф Фуркатович

Московский физико-технический институт
Кафедра интеллектуальных систем ФПМИ МФТИ
Научный руководитель: к.ф.-м.н. Бахтеев Олег Юрьевич

2025

Цель исследования

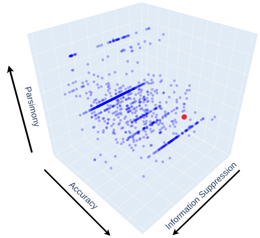
Проблема: Выбор оптимальной предсказательной модели сильно зависит от априорного знания человека о природе данных, т.е. от их индуктивного смещения. Определение индуктивного смещения автоматическим образом является открытой проблемой.

Примером индуктивного смещения могут служить свертки для изображений и временная зависимость для авторегресии.

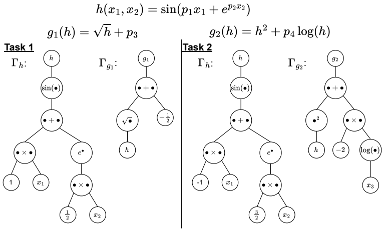
Цель: Предложить метод автоматического определения индуктивного смещения.

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

Архитектура решения



Модели с искомыми критериями.



Структура модели в виде дерева разбора.

Предлагается выбирать модель в режиме многозадачного обучения с учетом информационных критериев.

Гипотеза: модель, отражающая индуктивное смещение, располагается на Парето-фронте по следующим критериям: предсказательная способность, длина описания и взаимная информация между исходным объектом выборки и его представлением.

Постановка задачи

Пусть $\mathfrak{T} = \{T_i\}_{i=1}^m$ – множество задач. Задаче T_i соответствует набор данных $\mathfrak{D}_i = (\mathbf{X}^i, \mathbf{y}_i)$, где $\mathbf{X}^i \in \mathbb{R}^{N_i \times n}$, а $\mathbf{y}^i \in \mathcal{Y}$. Для регрессии $\mathcal{Y} = \mathbb{R}^{N_i}$ а для классификации $\mathcal{Y} = \{1, \dots, K\}^{N_i}$.

- Моделью $\mathbf{f}(\mathbf{x}; \mathbf{w})$ назовем отображение $\mathbb{R}^n \times \mathbb{R}^l \to \mathcal{Y}^m$, структура $\Gamma_{\mathbf{f}}$ модели которой представима в виде дерева разбора символьного выражения.
- Декомпозируем модель как $\mathbf{f} = \mathbf{g} \circ \mathbf{h} = (g_1 \circ \mathbf{h}, \dots, g_m \circ \mathbf{h})$. Модели $\mathbf{h}$ и $\mathbf{g}$ мы назовем *энкодером* и *декодером* соответственно.
- Назовем *индуктивным смещением* для декомпозируемой модели структуру первой части $\Gamma_{\mathbf{h}}$ модели $\mathbf{f}$.

Утверждение (Набиев, 2025)

Для любой модели $\mathbf{f} : \mathbb{R}^n \to \mathcal{Y}^m$ *существует такая модель* $\mathbf{h} : \mathbb{R}^n \to \mathbb{R}^d$ *и* $\mathbf{g} : \mathbb{R}^d \to \mathcal{Y}^m$, *что* $f = g \circ h$.

Постановка задачи

Теорема (Набиев, 2025)

Для скрытого пространства $\mathbb{R}^d$, такого что $d \geq m$, существует разложение $f = g \circ h$, где $g = \mathrm{Id}_Z$, а $Z \cong \mathcal{Y}^m$.

Энкодер $\mathbf{h}$ задает достаточную статистику относительно $\mathbf{y}$, если $\mathsf{I}(\mathbf{h}(\mathbf{X}), \mathbf{y}) = \mathsf{I}(\mathbf{X}, \mathbf{y})$.
Принцип Information Bottleneck (IB) — приближение минимальной достаточной статистики:

$$\min_{p(\mathbf{h}(\mathbf{x})|\mathbf{x})} \mathsf{I}(\mathbf{X}, \mathbf{h}(\mathbf{X})) - \beta\, \mathsf{I}(\mathbf{h}(\mathbf{X}), \mathbf{y}),$$

где параметр β задаёт баланс между сжатием и значимостью.
Для выполнения условия достаточности минимизируем $\mathsf{I}(\mathbf{X}, \mathbf{h}(\mathbf{X}))$:

$$\mathsf{I}(\mathbf{X}, \mathbf{h}(\mathbf{X})) - \beta\, \mathsf{I}(\mathbf{h}(\mathbf{X}), \mathbf{y}) \approx \mathsf{I}(\mathbf{X}, \mathbf{h}(\mathbf{X})) - \beta' \cdot L(f_i(\mathbf{X}), \mathbf{y}),$$

где L — функция потерь.

Постановка задачи

Задача оптимизации принимает вид:

$$\Gamma_{\mathbf{h}} = \underset{\mathbf{f}=(\mathbf{g}_1\circ\mathbf{h},\dots,\mathbf{g}_m\circ\mathbf{h})}{\arg\min} \frac{1}{m}\sum_{i=1}^{m} L_i(f_i(\mathbf{X}^i;\mathbf{w}_i^*),\mathbf{y}^i) + \lambda_1\, \mathsf{I}(\mathbf{X}^i,\mathbf{h}(\mathbf{X}^i)) + \lambda_2 C(\mathbf{f}),$$

$$\text{s.t. } \mathbf{w}_i^* = \underset{\mathbf{w}}{\arg\min}\, L_i(f_i(\mathbf{X}^i;\mathbf{w}),\mathbf{y}^i)$$

где $C : \mathfrak{F} \to \mathbb{R}$ — длина описания модели, с помощью кодирования Хаффмана.

Теорема (Набиев, 2025)

Пусть кол-во задач $m = 1$. Тогда существует решение f, которое можно разложить как $f = g \circ h$, где $g = \mathrm{Id}$, а $h = f$.

Следствие

Пусть имеется $m > 1$ задач, и пусть каждая задача T_i аппроксимируется функцией $f_i(\mathbf{x}) = [\mathbf{f}(\mathbf{x},\mathbf{w}_i)]_i$, где: $\Gamma_{g_i} = \Gamma_{g_j}$ для всех $i, j \in \{1,\dots,m\}$. Тогда существует декомпозиция вида $\mathbf{h}(\mathbf{x},\mathbf{w}) = \mathbf{f}(\mathbf{x};\mathbf{w}), \quad \mathbf{g}(\mathbf{x}) = \mathbf{x}$, такая что для каждой задачи $f_i(\mathbf{x}) = (\mathbf{g}\circ\mathbf{h})(\mathbf{x};\mathbf{w}_i) = \mathbf{h}(\mathbf{x};\mathbf{w}_i)$.

Данные для эксперимента

Эксперимент проводился на четырех типов выборок:

1. **Окружность**: 3 задачи: данные — точки окружности. Задачи имеют вид: $[r > 0], r + 0.5$ и $\sqrt{r} - 0.5$. Оптимальная модель — кривая второго порядка.
2. **Авторегрессия**: Задачи вида: $y_t = \alpha y_{t-1} + \varepsilon$. Модели оценивались по взаимной информации и длине описания.
3. **4-пикселя**: 8 задач: 6 бинарных классификаций и 2 регрессии. Оптимальная модель — использует локальную пространственную информацию.
4. **MNIST**: Сравнивались модели с масками, имитирующими свёртки, и случайными масками.

Гипотеза: структура энкодера будет содержать операции характерные для заданного типа выборки.

Результаты эксперимента: окружность и 4-пикселя

Окружность:

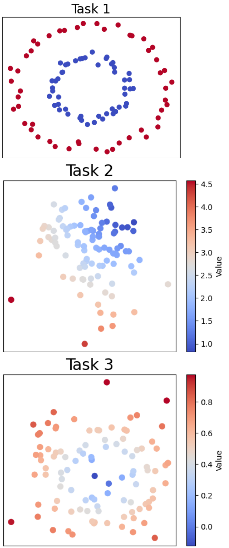
1. $h_0 = p_1$, $h_1 = x_0^2 + x_1^2$
2. $h_0 = (p_2 x_0 + x_1)^2$, $h_1 = p_1$
3. $h_0 = x_0^2 + x_1^2$, $h_1 = p_0$
4. $h_0 = p_2$, $h_1 = (x_1 + p_0 x_0)^2$
5. $h_0 = (x_0 + x_1)^2$, $h_1 = p_2$

Вывод: Модели 1, 3 и 5 демонстрируют высокую точность при минимальной сложности, соответствуя структуре задачи.

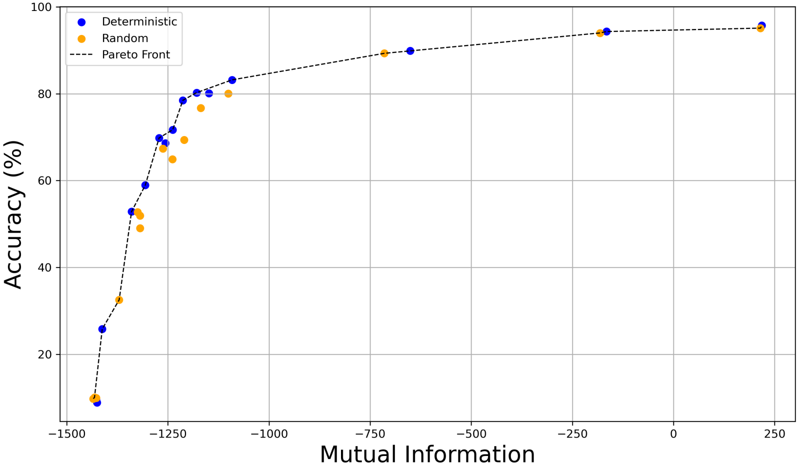
4-пикселя:

1. $h_0 = p_3 x_1 - x_0 - x_2 - p_5$, $h_1 = p_2 x_2 + x_3$
2. $h_0 = x_1$, $h_1 = \max(p_4, x_2 - x_0) + p_5 x_2 - x_3$
3. $h_0 = (x_1 + p_2)(p_3 x_3 + p_1 x_2)$, $h_1 = x_1 + p_1 x_0$
4. $h_0 = p_3 x_1 - x_3(p_0 + x_2)$, $h_1 = x_0 + p_2 x_2$
5. $h_0 = p_3(x_2 - p_4)(p_2 + x_0 + x_1)$, $h_1 = x_1(p_1 - x_3)$

Вывод: Модели 1 и 3 лучшие по компромиссу между точностью, длиной описания и локальностью признаков.



Результаты эксперимента: MNIST



Вывод: Модели учитывающие локальность достигают более высокой точности при заданной взаимной информации, т.е. они лежат на Парето-фронте.

Заключение

Выносится на защиту

1. Предложен критерий выбора модели для определения индуктивного смещения для заданого множества задач.
2. Приведены теоретические обоснования для предложенного метода.
3. Проведены эксперименты, которые показали, что модели с оптимальным соотношением качества, информации и сжатия располагаются на Парето-фронте.

- Результаты были доложены на 67-й конференции МФТИ.
- Планируется подача работы в рецензируемый журнал.