Выбор предсказательной модели в режиме многозадачного обучения с применением методов символьной регрессии

Набиев Мухаммадшариф Фуркатович
Научный руководитель: к.ф.-м.н. О. Ю. Бахтеев

Московский физико-технический институт
Кафедра интеллектуальных систем ФПМИ МФТИ

2024

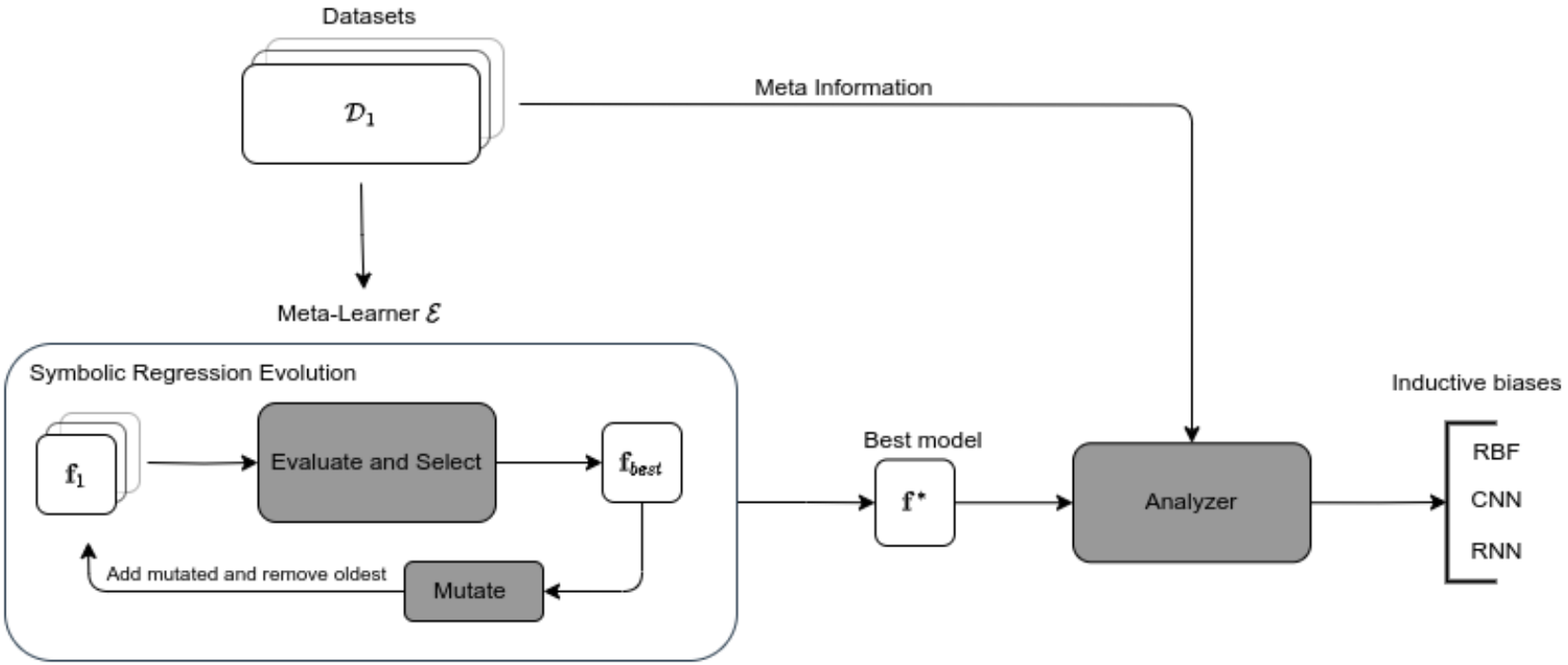
Цель исследования

Проблема: Построение архитектур моделей сильно зависит от априорного знания человека о природе данных, т.е. от их индуктивного смещения. Зная это выбирается соответствующий метод решения. Определение индуктивного смещения автоматическим образом является открытой проблемой.

Цель: Предложить метод автоматического извлечения индуктивного смещения.

Решение: Построение модели, решающей данную задачу, с помощью генетической символьной регрессии и извлечение индуктивного смещения из этой модели.

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

- Модель $\mathbf{f} \in \mathfrak{F}$ определяется набором из трех функций `Setup`, `Learn`, `Predict`.
- Мета-алгоритм $\mathcal{E} : \mathfrak{S} \to \mathfrak{F}$ представляет из себя генетический алгоритм, который конструирует модель путем символьной регрессии.
- Пусть $\mathrm{mACC}(\mathbf{f}, \mathfrak{S}) = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^{N_i} \frac{[f(\mathbf{x}_j)=\mathbf{y}_j]}{N_i}$. Тогда задача оптимизации сводится к нахождению наилучшей модели

$$\mathbf{f}^* = \arg\max_{\mathbf{f} \in \mathfrak{F}} \mathrm{mACC}(\mathbf{f}, \mathfrak{S}_{\text{test}}).$$

Данные для эксперимента

Для экспериментов использовались выборки `cricles` из `sklearn`. Выборки отличаются расположением центра концентрических кругов.

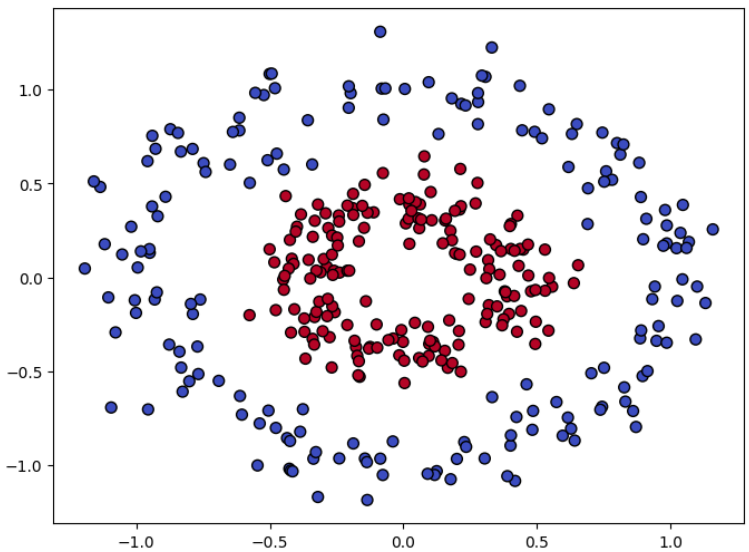


Рис.: Пример одной выборки.

- Количество выборок было равно 10. В каждом из них было по 100 элементов.
- Максимальная длина функций `Learn` и `Predict` была взята равной 10.

Гипотеза: код модели будет содержать элементы вычисления радиально-базисного ядра или схожих функций.

Результаты эксперимента

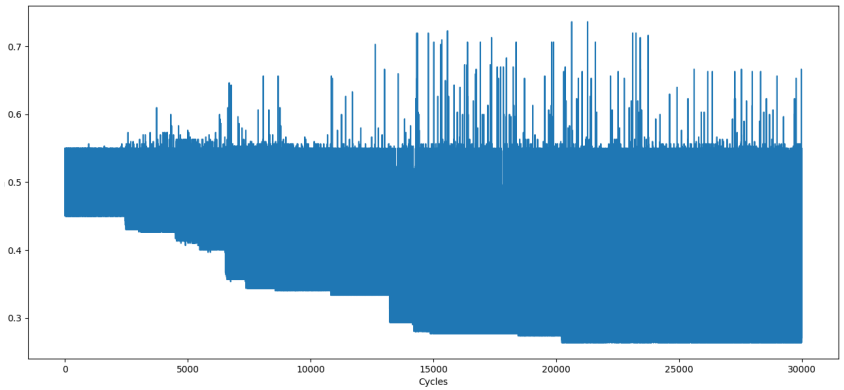


Рис.: График 1 – accuracy.

```
Predict:
  multiply(v0, v0)=v0
  exp(v0)=v1
  matmul(v0, v1)=s0
  subtraction(s3, s0)=s1
  matmul(v0, v3)=s2
  subtraction(s1, s2)=s1
  subtraction(s1, s0)=s1
```

Рис.: Функция `Predict` лучшей модели.

Формальная запись функции имеет вид

$$s_3 - 2(\mathbf{x} \odot \mathbf{x})^\intercal e^{\mathbf{x} \odot \mathbf{x}} - (\mathbf{x} \odot \mathbf{x})^\intercal \mathbf{v}_3,$$

где $\mathbf{v}_3$ и s_3 – веса модели. Данное представление близко к радиальным базисным функциям, которые являются индуктивным смещением данных.

Дальнейшие исследования

- Добавить анализатор для извлечения индуктивного смещения из модели.
- Добавление регуляризации учитывающей вложенность функций.
- Проведение экспериментов на выборках с индуктивным смещением CNN и RNN.
- Результаты будут доложены на конференции МФТИ.