УДК 004.023

Влияние метода выборки на анализ ландшафта приспособленности при настройке параметров эволюционных алгоритмов

Пикалов М.В. (Университет ИТМО)

В работе исследуется влияние метода и объема выборки на точность аппроксимации признаков ландшафта поиска (Exploratory Landscape Analysis, ELA) и на качество моделей машинного обучения, применяемых для настройки параметров эволюционных алгоритмов. В частности, на примере задачи регрессии параметров тестовой задачи W-model демонстрируется, что различные методы сэмплирования приводят к статистически значимо различным оценкам признаков, что критически важно для корректной работы прогнозных моделей.

Введение. Методы анализа ландшафта поиска предоставляют набор числовых признаков, описывающих свойства оптимизационных задач в условиях «черного ящика». Эти признаки используются моделями машинного обучения для автоматического выбора и настройки алгоритмов. Поскольку явная форма целевой функции неизвестна, признаки аппроксимируются по конечной выборке точек. Однако точность такой аппроксимации и ее зависимость от метода формирования выборки изучены недостаточно, что ставит под сомнение надежность подходов, основанных на ELA.

Основная часть. Для исследования влияния метода выборки была поставлена задача регрессии параметров (dummy, neutrality, epistasis, ruggedness) параметризованной тестовой задачи W-model по векторам из 35 признаков, вычисленных с помощью пакета flacco. Эксперименты проводились для пяти методов семплирования: вихрь Мерсенна, линейный конгруэнтный генератор RANDU, латинский гиперкуб (lhs), его оптимизированная версия (ilhs) и квазислучайные последовательности Фауре. Объем выборки варьировался: k = n, 10n, 100n. Для каждой конфигурации W-model генерировалось 1000 векторов признаков, на которых обучались и тестировались модели линейной регрессии (Ridge) и градиентного бустинга (Gradient Boosting).

Результаты показали, что увеличение объема выборки повышает устойчивость оценок признаков и снижает ошибку регрессии. Ключевым выводом является существенная зависимость аппроксимированных значений признаков от метода семплирования: распределения оценок для разных методов не совпадают даже при больших объемах данных. Наименьшая ошибка регрессии достигается при использовании последовательностей Фауре. Кросс-методологическое валидирование (обучение на выборках одного метода, тестирование на другом) подтвердило, что корректная работа модели возможна только при строгом совпадении стратегий выборки на этапах обучения и применения. Нарушение этого условия приводит к значительному росту ошибки предсказания.

Выводы. Аппроксимированные признаки ландшафта поиска не являются инвариантными относительно метода выборки. Их значения несут в себе «отпечаток» способа генерации точек, что требует обязательного использования единого метода семплирования на всех этапах работы с моделями машинного обучения. Полученные результаты указывают на перспективность применения квазислучайных последовательностей для повышения точности ELA. Для дальнейшего развития методов анализа ландшафта необходимо явно учитывать и изучать зависимость признаков от стратегии семплирования.

Пикалов М.В.