Меры сложности данных как инструмент анализа задач физико-химического моделирования

Вересников А.Е.

Автономная некоммерческая организация высшего образования

«Московский университет „Синергия“»

г. Москва, Российская Федерация

Методы машинного обучения находят применение в разных областях, включая моделирование физико-химических процессов, и в частности, солеотложения и коррозии нефтепромыслового оборудования под воздействием электромагнитных полей. Однако при решении подобных задач исследователь, как правило, располагает ограниченным объемом экспериментальных данных, что существенно затрудняет как выбор типа модели, так и настройку её гиперпараметров стандартными методами. В настоящей работе предлагается использовать меры сложности данных в качестве инструмента предварительного анализа.

Меры сложности данных представляют собой численные характеристики датасета, вычисляемые без привязки к конкретному алгоритму машинного обучения. Они исследуют и описывают внутреннюю структуру данных, приводя ее в численный вид. Ключевое преимущество подхода состоит в том, что эти характеристики вычисляются до начала обучения модели и позволяют составить формализованное представление о задаче ещё на этапе анализа данных.

Теоретическую основу подхода заложила работа Хо и Басу [1], получившая дальнейшее развитие в исследовании Лорены и соавторов [2], где число рассматриваемых мер сложности было доведено до двадцати двух с разбивкой на шесть категорий. Большинство этих мер изначально разрабатывалось для задач классификации, однако ряд из них, например, меры размерности и меры линейности, возможно использовать непосредственно в регрессионном контексте, тогда как адаптация части из них представляет собой самостоятельную исследовательскую задачу.

К ключевым категориям относятся следующие. Меры пересечения характеризуют способность отдельных признаков объяснять целевую переменную, в рассматриваемом контексте это позволяет оценить вклад каждого входного параметра в формирование наблюдаемого эффекта. Меры линейности определяют, допускает ли задача линейное описание, что напрямую влияет на выбор класса модели. Меры окрестности отражают локальную неоднозначность данных: насколько похожие по условиям эксперименты дают схожие результаты. Меры размерности характеризуют соотношение числа признаков и числа наблюдений, указывая на риск переобучения, особенно актуальный при малых выборках.

Вычисление мер сложности открывает два практически значимых направления. Первое связано с аналитикой и углубленным пониманием структуры задачи: характеристики датасета позволяют структурированно сравнивать экспериментальные группы, выявлять проблемные места в данных и интерпретировать их в предметных терминах. Кроме того, меры сложности могут служить основой для генерации синтетических данных: зная структурные характеристики реальных наблюдений, можно целенаправленно порождать датасеты с аналогичными свойствами и расширять обучающую выборку без проведения новых дорогостоящих экспериментов. Второе направление касается настройки гиперпараметров моделей. Стандартные методы оптимизации требуют многократного обучения модели и в условиях малых выборок оказываются либо ненадежными, либо вычислительно избыточными. Меры сложности, вычисленные до начала обучения, могут использоваться для предсказания эффективных значений гиперпараметров на основе опыта решения структурно близких задач. Такой подход относится к области мета-обучения и позволяет сократить вычислительные затраты на оптимизацию. В ранее проведенном исследовании он был апробирован на задаче классификации по следующей схеме. На 1890 синтетических датасетах с управляемой сложностью гиперпараметры оптимизировались байесовским поиском для трёх моделей: дерева решений, случайного леса и градиентного бустинга. Затем строился мета-

датасет, где признаками выступали меры сложности, а целевыми переменными были найденные оптимальные гиперпараметры. На нём обучались две регрессионные мета-модели, линейная регрессия и CatBoost, предсказывающие гиперпараметры по мерам сложности. Точность предсказания оказалась невысокой и зависела от глубины исходной оптимизации. При 70 итерациях метрика R^2 для ряда гиперпараметров была отрицательной. При увеличении до 700 итераций он рос, а при переходе с байесовской оптимизации на случайный поиск с 3000 итераций достигал 0,26 для глубины дерева решений. Практическая проверка, сравнение моделей с предсказанными гиперпараметрами против моделей с параметрами по умолчанию, дала неоднородный результат. Для дерева решений прирост F1 составлял до 0,005, тогда как для случайного леса и градиентного бустинга настроенные модели нередко уступали базовым. Эта неоднозначность, а также проверка исключительно на задаче классификации, делают проверку метода в регрессионном контексте отдельной и самостоятельно значимой задачей.

Практическая значимость подхода определяется спецификой предметной области. Даже незначительное повышение качества моделирования солеотложения и коррозии способно дать существенный экономический эффект: точное прогнозирование этих процессов позволяет сократить затраты на обслуживание и замену оборудования или снизить простои. В условиях, когда получение новых данных сопряжено с затратами, обоснованный выбор модели на основе имеющихся наблюдений приобретает самостоятельную практическую ценность.

Список использованной литературы

1. Ho T. K. Complexity measures of supervised classification problems / T. K. Ho, M. Basu // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2002. – Vol. 24, no. 3. – P. 289–300.

2. Lorena A. C. How complex is your classification problem? A survey on measuring classification complexity / A. C. Lorena, L. P. F. Garcia, J. Lehmann, M. C. P. Souto, T. K. Ho // ACM Computing Surveys. – 2019. – Vol. 52, no. 5. – P. 1–34.