УДК 004.82

ЭВОЛЮЦИОННЫЙ СИНТЕЗ ГИПЕРГРАФОВЫХ МОДЕЛЕЙ ДЛЯ ИНТЕРПРЕТИРУЕМОГО КОНСТРУИРОВАНИЯ ПРИЗНАКОВ ВЫСОКОГО ПОРЯДКА

Пшиншев Т. К.[1], Муравья Н. Р.[1]
Научный руководитель – канд. физ.-мат. наук, доцент Бойцев А. А.[1]
[1]Университет ИТМО
368706@niuitmo.ru

Введение

Современные информационные системы оперируют массивами данных высокой размерности и объема. Эффективность предиктивных моделей в подобных условиях в значительной степени определяется качеством используемого пространства признаков, что подтверждается фундаментальными работами в области машинного обучения [1]. Необходимость обеспечения интерпретируемости результатов в специализированных областях обуславливает вовлечение профильных экспертов для формирования физически обоснованных переменных. Следствием этого является высокая актуальность задач автоматического конструирования признаков (Automated Feature Engineering, AutoFE). Существующие алгоритмы на базе генетического программирования (GP) позволяют автоматизировать данный процесс, однако демонстрируют склонность к избыточному росту сложности расчетных схем (code bloat) и снижению прозрачности при моделировании взаимодействий высокого порядка [2]. В связи с этим разработка новых подходов к представлению расчетных структур, обеспечивающих баланс между выразительной способностью и интерпретируемостью итоговых моделей, является актуальной научно-технической задачей.

Основная часть

В рамках данного исследования предлагается концепция эволюционного синтеза гиперграфовых моделей (Hypergraph-Evolved Pipelines, HEP). В основе метода лежит использование функциональных гиперграфов – математических структур, где вершины представляют собой исходные признаки, а гиперребра выступают в роли функциональных отображений, преобразующих подмножества входных данных. В отличие от классических графов, гиперграфовое представление позволяет описывать k-арные взаимодействия признаков как единые логические блоки, обеспечивая высокую «представительскую плотность» модели. Это позволяет эффективно кодировать нелинейные зависимости высокого порядка без избыточного нарастания глубины расчетной схемы, что является критическим ограничением традиционных древовидных структур [3].

Процесс работы алгоритма HEP организован как итерационный цикл эволюционной оптимизации, что соотносится с современными парадигмами многокритериального поиска в системах автоматического машинного обучения (AutoML). На этапе инициализации формируется популяция кандидатов – наборов стохастически сгенерированных гиперребер с различными агрегирующими функциями. Оценка качества предложенных решений строится как композиция метрики точности решающей модели на преобразованных данных и регуляризационного штрафа за топологическую сложность гиперграфа. Такой комбинированный критерий приспособленности (фитнес-функция) обеспечивает автоматический отбор наиболее информативных и при этом лаконичных признаков.

Динамическое развитие архитектуры осуществляется через применение специализированных генетических операторов. Операторы топологической мутации

включают изменение состава связей внутри гиперребер, модификацию типов функций и структурное наращивание графа через генерацию новых взаимодействий. Для обмена эффективными признаковыми блоками в популяции применяется оператор макро-кроссовера, реализующий рекомбинацию подмножеств гиперребер между наиболее качественными особями.

Для обеспечения вычислительной эффективности системы используется механизм мемоизации на основе уникальных сигнатур топологии, исключающий повторные вычисления идентичных структур. Верификация метода проводится с применением методологии анализа чувствительности на специализированных наборах данных [4], имитирующих сложные нелинейные зависимости. Предварительные результаты показывают способность HEP выявлять скрытые закономерности высокого порядка при сохранении компактности итоговой структуры. В отличие от нейросетевых подходов, метод обеспечивает высокий уровень интерпретируемости: каждое сформированное гиперребро может быть напрямую декодировано в понятную аналитическую формулу, доступную для экспертной проверки.

Выводы

Результаты исследования подтверждают эффективность архитектуры функциональных гиперграфов при идентификации нелинейных связей высокого порядка. Адаптивность HEP к структурам данных различной природы позволяет синтезировать информативные признаки при сохранении компактности и прозрачности итоговых решений. Дальнейшие исследования будут направлены на интеграцию метода в открытые программные комплексы для структурного AutoFE и оптимизацию стратегий топологического поиска для повышения его эффективности в признаковых пространствах сверхвысокой размерности.

Литература

1. Domingos P. A few useful things to know about machine learning // Communications of the ACM. 2012. Vol. 55, № 10. P. 78–87.
2. Zhang X. et al. Automated Feature Engineering for AutoML Using Genetic Algorithms // Journal of Artificial Intelligence Research. 2024. Vol. 80. P. 543–572.
3. Miller J. F. Cartesian Genetic Programming. – Springer, 2011. 344 p.
4. Zakharov K., Boukhanovsky A. Model-Aware Automatic Benchmark Generation with Self-Error Instructions for Data-Driven Models // Machine Learning and Knowledge Extraction. 2025. Vol. 7, № 4. P. 148.