УДК 004.085

ГЕНЕРАЦИЯ НАБОРОВ ДАННЫХ ПО ЗАДАННЫМ МЕТА-ПРИЗНАКАМ С ПОМОЩЬЮ ЭВОЛЮЦИОННОГО ПРОГРАММИРОВАНИЯ

Галкин Г. Г.[1]
Научный руководитель – кандидат технических наук, доцент Забашта А.С.
[1]Университет ИТМО
galkin_gleb04@mail.ru

Введение

Исследование алгоритмов машинного обучения, включая задачи мета-обучения и автоматического машинного обучения, а также проведение соревнований, часто требует создания наборов данных с заданными характеристиками. Например, необходимы наборы данных, на которых один алгоритм демонстрирует низкую эффективность, а другой – высокую, или целые коллекции разнообразных наборов. Исторически, большое количество используемых алгоритмов синтеза данных опирались на генерацию данных с экстремальным значением качества работы алгоритма: наиболее трудные или простые случаи. Текущая ситуация характеризуется активным развитием методов генерации данных по мета-признаковому описанию, например с помощью генетического программирования [1], генеративно-состязательных сетей [2]. Недостатками существующих методов можно назвать плохую интерпретируемость результатов, низкую скорость работы, малое разнообразие возможных генерируемых данных.

Основная часть

Предлагаемый метод основан на эволюционном программировании для синтеза табличных наборов данных по заданному вектору мета-признаков. Ключевая идея заключается в поиске не самого набора данных с заданными признаками, а порождающей его функции. Это позволяет вместе с требуемым набором данных получить описание зависимостей в этом наборе. Генератор строит набор символьных выражений, которые порождают набор данных. В качестве поискового оператора используется генетическое программирование – кроссовер и мутация. На каждой итерации алгоритма вычисляются признаки, формируется таблица, оцениваются мета-признаки и расстояние до целевого вектора – fitness-функция. По fitness-функции отбираются родители и применяются операторы поиска. Такой подход позволяет получить набор данных вместе с функцией, отражающей зависимости в сгенерированном наборе данных. Благодаря тому, что в ходе алгоритма изменяется функция, а потом из нее генерируется набор данных, предлагаемый метод может уменьшить время генерации в сравнении с существующими методами. Более того, алгоритм дает большую интерпретируемость результатов за счет предоставления функции генерации.

Выводы

Реализован описываемый метод генерации табличных наборов данных, основанный на генетическом программировании выражений для признаков, проанализированы полученные результаты, проведено сравнение с существующими решениями.

Литература

1. Забашта А.С., Фильченков А.А. Построения наборов данных для задачи бинарной классификации по их характеристическому описанию // Научно-технический вестник информационных технологий, механики и оптики. 2017. №3.

2. Качальский И.В., Забашта А.С. (науч. рук. Фильченков А.А.) Использование порождающих состязательных сетей для синтеза наборов данных в матричном представлении // Сборник тезисов докладов конгресса молодых ученых. Электронное издание. – СПб: Университет ИТМО, 2019. Режим доступа: https://kmu.itmo.ru/digests/article/221 (Дата обращения 15.02.2026).