УДК 004.85

ГЕНЕРАЦИЯ ДАННЫХ ДЛЯ ПОВЫШЕНИЯ КАЧЕСТВА МЕТА-КЛАССИФИКАЦИИ

Забашта А.С. Университет ИТМО
Научный руководитель – к.ф.-м.н., Фильченков А.А.
Университет ИТМО

В данной работе рассматриваются различные методы генерации данных с заданными характеристиками для задач машинного обучения. Один из подходов основан на эволюционных алгоритмах, а второй на основе генеративно-состязательных сетей.

Введение. В области мета-обучения алгоритмы машинного обучения используются для анализа самих алгоритмов машинного обучения. Для подобного анализа требуется большая коллекция разнообразных наборов данных. Существующих наборов данных не достаточно, а если произвольно генерировать наборы данных, то их характеристики будут смещены в сторону генератора и будут отличаться от характеристик реальных наборов данных.

Основная часть. Данную проблему можно решить двумя способами: сведением к задаче минимизации и используя подход на основе генеративно-состязательных сетей.

В первом способе используется функция ошибки – расстояние в мета-признаковом пространстве между характеристиками полученного набора данных и требуемыми характеристиками. Данную функцию предлагается минимизировать, используя эволюционные алгоритмы. Для применения эволюционных алгоритмов требуется разработать операторы кроссовера и мутации. Помимо наивного представления набора данных в виде матрицы чисел, которая разворачивается в вектор, предлагается использовать естественные преобразования: удаление и добавление атрибутов и объектов в набор данных. Такое преобразование не учитывает порядок строк и столбцов в наборе данных, что облегчает поиск в полученном пространстве.

Для применения генеративно-состязательных сетей требуется построить две дифференцируемых функции: генератор – отображающий вектор характеристик в набор данных и дискриминатор – отображающий набор данных в метку класса. В качестве метки класса может использоваться информация об алгоритме машинного обучения наиболее оптимальном для данного набора данных. Проблема данного подхода в том, что существующие архитектуры сетей рассчитаны на обработку изображений, в которых важен порядок строк и столбцов в советующей матрице. Данную проблему можно решить, если в архитектуре генеративно-состязательных сетей между генератором и дискриминатором поставить вспомогательную функцию приводящую набор данных к общему виду путём перестановки его строк и столбцов.

Выводы. Предложенные методы генерации наборов данных с требуемыми характеристиками для задач машинного обучения можно использовать для улучшения качества систем мета-классификации в сценарии активного обучения, если в качестве функции ошибки взять неуверенность классификатора, либо использовать методы повышения разнообразия мета-признаковых пространств. Также их можно использовать для исследования самих алгоритмов, если пытаться сгенерировать набор данных наиболее сложный для одного алгоритма и в тоже время наиболее простой для другого.