УДК 004.8

АЛГОРИТМ ВЫБОРА ПРИЗНАКОВ ДЛЯ ЧАСТИЧНО-РАЗМЕЧЕННЫХ НАБОРОВ ДАННЫХ НА ОСНОВЕ ГРАФОВЫХ ПОДХОДОВ

Макеев П.А. (Университет ИТМО, Факультет Информационных Технологий и Программирования)

Научный руководитель – кандидат технических наук Сметанников И.Б.
(Университет ИТМО, Факультет Информационных Технологий и Программирования)

В докладе рассматривается алгоритм выбора признаков для частично-размеченных наборов данных на основе графовых подходов.

В современном машинном обучении и анализе данных задачи обработка исходного набора данных являются особенно важной в силу того, что часто исходные данные имеют высокую размерность и страдают от нескольких проблем, связанных с этим, которые объединяются в общую формулировку «проклятие размерности». Среди них низкое качество работы обученных на таких данных моделей, длительность обучения, из-за шумовых признаков и признаков, плохо отражающих зависимость с целевым значением. Чтобы исправить подобные проблемы разрабатываются алгоритмы выбора и конструирования признаков, уменьшающие размерность исходного набора данных, до нескольких, наиболее точно отражающих зависимость между данными и метками классов. Существует несколько классических категорий подходов. Фильтрующие алгоритмы – отбирающие признаки на основе значений метрик корреляция признаков, они обладают высокой скоростью работы, но часто неточно отбирают необходимое множество признаков. Алгоритмы обертки – основанные на поиске нужного множества признаков, путем обучения модели и оценки качества набора на мерах работы модели, такие методы позволяют точно подобрать набор признаков, но склонны к переобучению. Ансамблирующие алгоритмы, строящиеся на комбинациях других видов алгоритмов, часто обладающие достаточно сложной структурой оценки признаков**.**

На основе статьи планируется реализовать алгоритм выбора признаков для расширения функционала библиотеки ITMO_FS. Основной идеей выбора признаков для частично-размеченных наборов данных на основе графовых подходов является минимизация лосс-функции и регуляризации: для устойчивости к выбросам и для учета связей между признаками. Алгоритм также характерен тем, что учитывает два вида данных – размеченные и не размеченные. Такой подход раннее не рассматривался в библиотеке, что значительно расширит ее инструментарий и позволит проще решать более широкий спектр задач.

Макеев П.А. (автор) Подпись

Сметанников И.Б. (научный руководитель) Подпись