УДК 004.2

ГИБРИДНЫЙ АНСАМБЛИРУЮЩИЙ АЛГОРИТМ ВЫБОРА ПРИЗНАКОВ, СТРОЯЩИЙСЯ НА КОМБИНАЦИИ ВЕСОВЫХ ФУНКЦИЙ ЗНАЧИМОСТИ ПРИЗНАКОВ РЕКУРСИВНО-ИСКЛЮЧАЮЩИХ ОБЕРТОК НАД МЕТОДОМ ОПОРНЫХ ВЕКТОРОВ, СЛУЧАЙНЫХ ЛЕСОВ И ГРАДИЕНТНОГО БУСТИНГА

Балашов Я.Д. (Университет ИТМО, Факультет Информационных Технологий и Программирования),
Научный руководитель – к.т.н., доцент Сметанников И.Б.
(Университет ИТМО, Факультет Информационных Технологий и Программирования)

В наше время сильно актуальна задача понижения размерности. Поэтому в данной работе представлены реализации алгоритмов выбора признака на основе техники рекурсивного исключения, а также комбинации этих алгоритмов.

Введение. Задача понижения размерности крайне актуальна для задач машинного обучения и анализа данных по ряду причин: проклятие размерности – большое количество признаков негативно влияет на скорость обучения и предсказания моделей, а также на качество решения поставленной задачи; разреженность некоторых данных и пропущенные значения; шумовые признаки, не вносящие полезный вклад в решение задачи.
Для борьбы с подобными проблемами были созданы алгоритмы понижения размерности. Существует два типа алгоритмов понижения размерности – алгоритмы выбора признаков и алгоритмы комбинирования признаков. Наиболее часто применяются алгоритмы выбора признаков, их существует несколько видов: фильтрующие алгоритмы – для каждого признака считается мера зависимости целевого значения от него, а затем при помощи отсекающего правила отбираются наиболее важные; алгоритмы обертки – отбираются признаки с помощью некоторого алгоритма, затем обучают модель на этих данных и при помощи метрик оценивают качество работы полученной модели, после чего исходный алгоритм отбора оптимизируется; ансамблирующие и гибридные алгоритмы – используют различные комбинации других видов алгоритмов понижения размерности; встроенные алгоритмы – спроектированные под конкретную задачу.
Поскольку описанных алгоритмов выбора признаков достаточно много и каждый из них имеет свои преимущества на разных задачах выбора признаков библиотеку ITMO_FS необходимо постоянно расширять и добавлять новые алгоритмы и методы. Гибридные алгоритмы, упомянутые выше, слабо представлены среди множества алгоритмов, уже реализованных в библиотеке ITMO_FS. Поэтому целью данной работы является реализация алгоритма выбора признаков по подходу, описанному в некоторой статье, с использованием инструментов библиотеки ITMO_FS для последующего добавления в нее.

Основная часть.
На старте алгоритма обучается модель, во время обучения которой будут найдены веса признаков - они отражают их важность. После этого все признаки ранжируются в соответствии с весами и самый низкий по рангу удаляется, и так до тех пор, пока множество признаков не будет пусто, затем на основе полученного списка рангов отбираются признаки.
Для метода опорных векторов похожий подход измеряет важность признаков с помощью ядерных функций, которые используются для как можно лучшего разделения высокоразмерных данных. Обычно в таких случаях применяются линейные ядерные функции, чьи веса можно использовать как метрику важности признаков.
Для случайного леса метрики важности признаков рассчитываются на основе либо среднего ухудшения точности, которая показывает, как меняется точность модели в зависимости от набора признаков, либо среднего ухудшения Gini, которая показывает вклад признака в однородность вершин в дереве.

В случае градиентного бустинга важность признака оценивается при помощи индекса Гини. Предлагается использовать ансамбль этих трех методов, путем подсчета важности всех признаков каждым алгоритмом выбора признаков и последующей агрегацией этих метрик. Основные два варианта — это сумма значений метрик или их линейная комбинация с весами – точностью работы моделей. Для лучшей работы ансамбля рекомендуется использовать нормализацию весов.

Выводы. В рамках данной работы были реализованы и добавлены в библиотеку ITMO_FS несколько алгоритмов выбора признаков на основе техники рекурсивного исключения, а также способа комбинации этих алгоритмов.

Балашов Я.Д. (автор) Подпись

Сметанников И.Б. (научный руководитель) Подпись