Секция «Теория вероятностей и математическая статистика»

Анализ эпидемиологических данных с помощью случайного леса

Научный руководитель – Яровая Елена Борисовна

Перевердиева Ксения Геннадьевна
Студент (специалист)
Московский государственный университет имени М.В.Ломоносова,
Механико-математический факультет, Кафедра теории вероятностей, Москва, Россия
E-mail: pereverdieva@gmail.com

Алгоритм машинного обучения, называемый деревом решений позволяет решать как задачи классификации, так и задачи регрессии и является альтернативой регрессионным моделям [9,10]. Алгоритм "случайный лес"это модификация стандартных деревьев решений, которая менее чувствительна к выбору параметров и разделению данных на обучающую и тестовую выборки [1,10]. Данный алгоритм и его модификации широко используются при решении прикладных задач, таких как распознавание объектов, классификация текстов, распознавание жестов, обнаружение спама, обучение ранжированию в информационном поиске [2-4,10]. Однако случайный лес не получил широкого применения в эпидемиологических исследованиях, так как нет однозначной интерпретации результатов и требует для эффективного использования больших вычислительных мощностей [5,10]. Большинство работ, основанных на эпидемиологических данных, используют стандартные регрессионные модели для анализа сердечно сосудистых заболеваний [6-8]. Преимущество случайного леса выражается в возможности моделировать сложные нелинейные зависимости, в отличие от линейной и логистической регрессий.

Целью работы было изучение особенностей применения алгоритма "случайный лес"к эпидемиологическим данным и его сравнение с логистической регрессией. Кроме того, была задача изучить взаимосвязь между коэффициентами регрессии и, так называемыми, важностями признаков, определяемыми для случайного леса, для получения более полной картины о степени влияния факторов риска. В результате анализа выявлена статистически значимая, но относительно небольшая разница в предсказательных способностях алгоритмов. AUC ROC для случайного леса составил 81.56%, для логистической регрессии с непрерывными переменными - 82.27%, для логистической регрессии с бинарными переменными - 81.12%. Таким образом, показано, что предположение о наличии линейных связей при анализе гипертонии является оправданным. Анализ данных произведен с помощью языка программирования Python 3. Данные предоставлены отделом эпидемиологии хронических неинфекционных заболеваний ФГБУ НМИЦ ТПМ Минздрава РФ.

Таким образом, показано, что предположение о наличии линейных связей при анализе гипертонии является оправданным. Также можно сказать, что стандартные регрессионные модели хорошо моделируют эпидемиологические данные, при этом позволяя использовать статистические методы, такие как проверка гипотез. С другой стороны, случайный лес позволяет выявлять нелинейные зависимости и имеет более интерпретируемые важности признаков, чем коэффициенты регрессии.

Источники и литература

1) А.Г. Дьяконов «Методы решение задач классификации с категориальными признаками»
2) С.П.Чистяков «Случайные леса: обзор»

3) Д.В. Дресвянский «Эффективность методов интеллектуального анализа данных при распознавании спама»

4) И.С. Веретенников, Е.А. Карташев «Оценка качества классификации текстовых материалов с использованием алгоритма машинного обучения «случайный лес»»

5) И.Л.Кафтанников, А.В. Парасич «Особенности применения деревьев решения в задачах классификации»

6) А.М. Ерина, О.П. Ротарь и др. «Предгипертензия и кардиометаболические факторы риска (по материалам исследования ЭССЕ-РФ)»

7) С. А. Максимов, А. Е. Скрипченко и др. «Связь курения с ишемической болезнью сердца и факторами сердечно-сосудистого риска (исследование ЭССЕ-РФ в Кемеровской области)»

8) С.А. Максимов, Д.П. Цыганкова «Популяционный риск развития ишемической болезни сердца в зависимости от объемов потребления алкоголя населением (исследование ЭССЕ-РФ в Кемеровской области)»

9) L Breiman, JH Friedman «Classification algorithms and regression trees»

10) Курс "Обучение на размеченных данных"на сайте www.coursera.org