Исследование и разработка древовидных моделей для задачи анализа выживаемости

Васильев Юлий Алексеевич

Студент, 2 курс магистратуры

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: iuliivasilev@gmail.com

Научный руководитель *— Петровский Михаил Игоревич*

Анализ выживаемости является актуальной областью исследований в современном обществе. Основная задача данной области – предсказать вероятность наступления события и времени до него. В такой формулировке область имеет множество сфер применения: здравоохранение, кредитный скоринг и инженерные науки. В инженерных науках анализ выживаемости используется для решения задач анализа надежности, предполагая в качестве события сбой или отказ системы. В сфере здравоохранения в качестве события может быть выбран исход по итогу лечения пациента: смерть, рецидив, выздоровление.

Статистические методы требуют большого количества данных, однако полные данные могут быть недоступны. В неполных данных время до наступления события может быть неизвестно по нескольким причинам (например, событие не наступило до окончания событий или наблюдение покинуло исследование). В таком случае, события называют цензурированными.

Фундаментальная задача анализа выживаемости может быть сведена к оценке функции выживания, определяющую вероятность того, что событие наступит позже определенного времени.

Наиболее популярным методом в области анализа выживаемости является метод пропорциональных рисков Кокса [1]. Однако, данный метод имеет ряд недосатков: строгие предположения, сложность зависимости времени события от предикторов, работа только на заполненных данных, независимость значимости признаков от времени.

Для избежания недостатков применяются древовидные модели машинного обучения [2], основанные на рекурсивном разбиении выборки на подвыборки с различающейся выживаемостью. Для определения различия между выборками используются log-rank критерий, учитывающий цензурирование данных.

Однако, существующие методы не обрабатывают пропуски в данных, а log-rank критерий обладает малой чувствительностью к особенностям наборов данных. В частности, важность ранних событий

учитывается взвешенными log-rank критериями [3]: wilcoxon, taroneware, peto.

В данной работе были разработаны древовидные модели, основанные на взвешенных log-rank критериях и учитывающие пропуски в данных. Также разработаны беггинг и бустинг ансамбли деревьев решений. Данный подход позволяет работать с пропущенными данными, имеет большую чувительность с входным данным, а использование поправки Бонферрони для выбора лучшего признака при разбиении позволяет корректнее сравнивать признаки с разным количеством значимых разбиений.

Экспериментальное исследование и оценка качества была проведена на международных наборах данных PBC [4], GBSG [5] и Wuhan [6]. Были рассмотрены метрики: concordance index, integrated brier score, integrated AUC.

В ходе проведённых экспериментальных исследований, предложенные алгоритмы позволили добиться лучшего результата по сравнению с классически методами. Предложенные алгоритмы были реализованы в виде open-source библиотеки на языке Python 3.8.

Литература

1. Cox D. R. Regression models and life-tables //Journal of the Royal Statistical Society: Series B (Methodological). – 1972. – Т. 34. – №. 2. – С. 187-202.
2. Fernandez C. et al. Experimental Comparison of Semi-parametric, Parametric, and Machine Learning Models for Time-to-Event Analysis Through the Concordance Index //arXiv preprint arXiv:2003.08820. – 2020.
3. Lee S. H. Weighted Log-Rank Statistics for Accelerated Failure Time Model //Stats. – 2021. – Т. 4. – №. 2. – С. 348-358.
4. Kaplan M. M. Primary biliary cirrhosis //New England Journal of Medicine. – 1996. – Т. 335. – №. 21. – С. 1570-1580.
5. Sauerbrei W., Royston P. Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials //Journal of the Royal Statistical Society: Series A (Statistics in Society). – 1999. – Т. 162. – №. 1. – С. 71-94.
6. Yan L. et al. An interpretable mortality prediction model for COVID-19 patients //Nature machine intelligence. – 2020. – Т. 2. – №. 5. – С. 283-288.