

DOI: 10.37791/2687-0649-2025-20-6-105-120

Обобщенная модель внимания для анализа выживаемости сложных объектов при цензурированных данных

Л. В. Уткин¹, А. В. Константинов¹, Н. М. Вербова¹, И. В. Ильин^{1*}

¹Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, Россия
*igor.ilin@spbstu.ru

Аннотация. Самые различные прикладные области, включая медицину, безопасность, экономику, промышленность, связаны с моделированием процессов наступления различных событий, таких как выздоровление пациента, финансовое банкротство предприятия, отказ промышленного оборудования и т. д. Моделирование этих процессов может выполняться в рамках анализа выживаемости (survival analysis), являющегося статистическим методом анализа времени до событий и отличительной особенностью которого от многих других статистических методов и методов машинного обучения является наличие цензурированных данных, имеющих место тогда, когда событие не наблюдается и известно только то, что оно не произошло до определенного момента времени. Цензурированные данные значительно усложняют моделирование и предсказание наступления критических событий. Одним из эффективных инструментов анализа выживаемости при наличии цензурированных данных является машинное обучение. В частности, современные модели машинного обучения на основе трансформеров демонстрируют перспективные результаты в анализе выживаемости благодаря своей способности учитывать сложные зависимости. Однако стандартный механизм внимания в них часто игнорирует фундаментальную структуру данных времен до наступления события, а именно различие между цензурированными и нецензурированными наблюдениями. Для преодоления этого недостатка в работе предлагается новая модель и новый подход к реализации механизма внимания, который переопределяет веса внимания за счет включения априорных характеристик анализа выживаемости на основе оценки Берана или модели Кокса. Вместо того чтобы полагаться исключительно на расстояния между векторными представлениями признаков, как это осуществляется в современных моделях, предлагаемая модель вычисляет веса внимания как взвешенную линейную комбинацию компонентов, полученных из основных априорных характеристик анализа выживаемости, таких как расстояния между оценками функций выживаемости или математическими ожиданиями времени до наступления события для различных обучающих объектов. Предлагаемый подход позволяет существенно расширить класс трансформенных моделей анализа выживаемости для получения более высокой точности предсказаний. Алгоритм, реализующий предлагаемую модель, является основой для трансформеров. Эксперименты на реальных наборах данных подтверждают, что обобщенная модель обеспечивает наилучшее предсказание среди ряда известных моделей.

Ключевые слова: анализ выживаемости, машинное обучение, модель внимания, оценка Берана, трансформер, цензурированные данные

Для цитирования: Уткин Л. В., Константинов А. В., Вербова Н. М., Ильин И. В. Обобщенная модель внимания для анализа выживаемости сложных объектов при цензурированных данных // Прикладная информатика. 2025. Т. 20. № 6. С. 105–120. DOI: 10.37791/2687-0649-2025-20-6-105-120

© Уткин Л. В., Константинов А. В., Вербова Н. М., Ильин И. В., 2025.

A generalized attention model for survival analysis of complex objects under censored data

L. Utkin¹, A. Konstantinov¹, N. Verbova¹, I. Ilin^{1*}

¹Peter the Great St. Petersburg Polytechnic University, Saint Petersburg, Russia
igor.ilin@spbstu.ru

Abstract. A wide variety of applied fields, including medicine, security, economics, and industry, are concerned with modeling the processes of various events occurring, such as a patient's recovery, a company's financial bankruptcy, industrial equipment failure, etc. Their modeling can be performed within the framework of survival analysis, a statistical method for analyzing time-to-event data whose distinctive feature, setting it apart from many other statistical and machine learning methods, is the presence of censored data. This occurs when an event is not observed and it is only known that it did not happen before a certain point in time. Censored data significantly complicates the modeling and prediction of critical events. Machine learning is an effective tool for survival analysis in the presence of censored data. In particular, modern transformer-based machine learning models demonstrate promising results in survival analysis due to their ability to account for complex dependencies. However, the standard attention mechanism in these models often ignores the fundamental structure of time-to-event data, namely, the distinction between censored and uncensored observations. To overcome this shortcoming, this paper proposes a new model and a new approach to implementing an attention mechanism that redefines attention weights by incorporating prior characteristics of survival analysis based on the Beran estimator or the Cox model. Instead of relying solely on distances between feature vector representations, as is done in current models, the proposed model computes attention weights as a weighted linear combination of components derived from key prior characteristics of survival analysis, such as distances between survival function estimates or time-to-event expectations for different training objects. The proposed approach enables a significant expansion of the class of transform models for survival analysis, achieving higher prediction accuracy. The algorithm implementing the proposed model is the basis for transformers. Experiments on real datasets confirm that the generalized model provides the best prediction among a number of known models.

Keywords: survival analysis, machine learning, attention model, Beran estimator, transformer, censored data

For citation: Utkin L., Konstantinov A., Verbova N., Ilin I. A generalized attention model for survival analysis of complex objects under censored data. *Prikladnaya informatika*=Journal of Applied Informatics, 2025, vol.20, no.6, pp.105-120 (in Russian). DOI: 10.37791/2687-0649-2025-20-6-105-120

© Utkin L., Konstantinov A., Verbova N., Ilin I., 2025.

Введение

Цензурированные данные и анализ выживаемости является критически важным инструментом в статистике, используемым для моделирования данных о времени до наступления собы-

тия в различных областях, включая здравоохранение, технику и финансы. Так, в здравоохранении анализ выживаемости позволяет прогнозировать критические исходы, такие как ремиссия заболевания и восстановление, смерть пациента, что,