2024, **25** (3), 357–377. doi 10.26089/NumMet.v25r328

doi 10.26089/NumMet.v25r328 УДК 004.852

Применение регуляризации при вычислении критериев разбиения в моделях анализа выживаемости

Ю. А. Васильев
Московский государственный университет имени М. В. Ломоносова,
факультет вычислительной математики и кибернетики,
кафедра интеллектуальных информационных технологий,
Москва, Российская Федерация
ORCID: 0000-0001-9210-5544, e-mail: iuliivasilev@gmail.com

М. И. Петровский
Московский государственный университет имени М. В. Ломоносова,
факультет вычислительной математики и кибернетики,
кафедра интеллектуальных информационных технологий,
Москва, Российская Федерация
ORCID: 0000-0002-1236-398X, e-mail: michael@cs.msu.su

И. В. Машечкин
Московский государственный университет имени М. В. Ломоносова,
факультет вычислительной математики и кибернетики,
кафедра интеллектуальных информационных технологий,
Москва, Российская Федерация
ORCID: 0000-0002-9837-585X, e-mail: mash@cs.msu.su

Аннотация: Методы анализа выживаемости решают задачу описания и прогнозирования событий. Модели учитывают случаи цензурирования, в которых истинное время события неизвестно из-за выхода наблюдения из исследования. Статистические методы предполагают, что цензурирование неинформативно и связь между причиной выхода наблюдения и проведением исследования отсутствует. В работе проводится исследование влияния информативности на эффективность статистических методов. В частности, критерий log-rank используется для сравнения функций риска и имеет низкую чувствительность в случае малых выборок или мультимодального распределения времени события. Для преодоления недостатков предлагается метод вычисления регуляризованных критериев, которые используют информацию об априорном распределении событий во времени и оценивают различия между функциями риска для всех моментов времени. Метод регуляризации был интегрирован в метод построения деревьев выживания и привел к улучшению качества прогнозирования на четырех медицинских наборах данных. Кроме того, предложенный метод превзошел существующие статистические методы и реализацию дерева выживания на всех наборах данных.

Ключевые слова: анализ выживаемости, информативность цензурирования, критерии разбиения, регуляризация.

Для цитирования: Васильев Ю.А., Петровский М.И., Машечкин И.В. Применение регуляризации при вычислении критериев разбиения в моделях анализа выживаемости // Вычислительные методы и программирование. 2024. **25**, № 3. 357–377. doi 10.26089/NumMet.v25r328.

2024, **25** (3), 357–377. doi 10.26089/NumMet.v25r328

Applying regularization to calculate split criterion for survival models

Iulii A. Vasilev
Lomonosov Moscow State University,
Faculty of Computational Mathematics and Cybernetics,
Department of Intelligent Information Technologies, Moscow, Russia
ORCID: 0000-0001-9210-5544, e-mail: iuliivasilev@gmail.com

Mikhail I. Petrovskiy
Lomonosov Moscow State University,
Faculty of Computational Mathematics and Cybernetics,
Department of Intelligent Information Technologies, Moscow, Russia
ORCID: 0000-0002-1236-398X, e-mail: michael@cs.msu.su

Igor V. Mashechkin
Lomonosov Moscow State University,
Faculty of Computational Mathematics and Cybernetics,
Department of Intelligent Information Technologies, Moscow, Russia
ORCID: 0000-0002-9837-585X, e-mail: mash@cs.msu.su

Abstract: Survival analysis methods solve the problem of describing and predicting events. Models account for cases of censoring in which the true time of the event is unknown due to the withdrawal of the observation from the study. Statistical methods assume that censoring is uninformative and there is no relationship between the reason for the observation withdrawal and the study. This paper investigates the effect of informativeness on the performance of statistical methods. In particular, the log-rank criterion is used to compare hazard functions and has low sensitivity in the case of small samples or multimodal event time distribution. To overcome the shortcomings, we propose a method to compute regularized criteria that use a priori information about the distribution of events over time and evaluate the differences between risk functions for all time points. The regularization method was integrated into the survival tree method and resulted in improved prediction quality on four medical datasets. Also, the proposed method outperformed the existing statistical methods and survival tree realization on all datasets.

Keywords: survival analysis, informative censoring, splitting criteria, regularization.

For citation: I. A. Vasilev, M. I. Petrovskiy, I. V. Mashechkin, "Applying regularization to calculate split criterion for survival models," Numerical Methods and Programming. **25** (3), 357–377 (2024). doi 10.26089/NumMet.v25r328.

1. Введение. Анализ выживаемости является мощным инструментом для описания и прогнозирования событий. Для сбора данных определяется интервал исследования, в рамках которого фиксируются признаки и исход наблюдений. На момент входа в исследование наблюдению сопоставляется вектор признаков, который может обогащаться при повторном измерении значений. Модели выживаемости позволяют решать множество прикладных задач: оценка влияния признаков, прогноз вероятности и времени наступления события, определение статистически значимых различий в выживаемости двух или более групп.

На практике полные данные могут быть недоступны из-за ограниченности контроля за наблюдениями или наличия временны́х рамок. В неполных данных время до наступления события может быть неизвестно по нескольким причинам (например, выход из исследования по желанию пациента или потеря наблюдения). Наблюдения с известным истинным временем называются терминальными, а с неопределенным временем — цензурированными. Наиболее распространенным является правое цензурирование, при котором известно время выхода из исследования до наступления определенного события. Таким образом,

2024, **25** (3), 357–377. doi 10.26089/NumMet.v25r328

модели выживаемости учитывают две целевые переменные: время события T и флаг цензурирования δ, а также вектор признаков X наблюдения при входе в исследование. Это делает модели выживаемости более гибкими и точными по сравнению с другими статистическими методами.

Уникальной особенностью анализа выживаемости является возможность прогнозирования вероятности наступления события во времени. Функция выживания (survival function) определяет вероятность ненаступления события по истечении определенного времени $S(t) = P(T \geqslant t)$, где t — время наблюдения, T — случайная величина времени события. Функция плотности смертности (death density function) $f(t) = \big(1 - S(t)\big)'$ определяет риск наступления события в конкретный момент времени, а функция риска (hazard function) $h(t) = f(t)/S(t)$ определяет относительный риск наступления события в момент t при условии, что событие не наступило ранее.

Наиболее распространенные статистические методы основаны на строгих теоретических предположениях, которые могут нарушаться на практике и приводить к смещению прогнозов моделей. В частности, статистические методы используют предположение о неинформативном цензурировании [1–3]. Цензурирование называется неинформативным, если причины цензурирования не связаны с проведением исследования, и информативным, если причины связаны с неучтенными факторами исследования. Стоит отметить, что информативность цензурирования приводит к появлению выборок с мультимодальным распределением времени события.

Целью данной работы является анализ влияния информативного цензурирования на качество и интерпретируемость деревьев выживания. В отличие от статистических моделей, деревья выживания не имеют строгих предположений и строят интерпретируемые прогнозы на основе правил разбиения выборки. В ходе исследования мы рассматриваем свойства и недостатки получаемого прогноза и предлагаем новый метод построения деревьев выживания с регуляризованным критерием разбиения, применимый к случаям информативного цензурирования в данных.

Статья организована следующим образом. В разделе 2 представлен обзор используемых наборов медицинских данных, статистических моделей и метрик качества. В разделе 3 рассмотрена чувствительность деревьев выживания к малым выборкам и мультимодальным распределениям времени событий и предлагается новый метод регуляризации критериев log-rank. В разделе 4 предложена модификация метода построения деревьев выживания, которая преодолевает недостатки мультимодального распределения времени событий. В разделе 5 приведены результаты экспериментального исследования влияния предложенных модификаций на метрики качества и проведен анализ предложенных и существующих методов. В разделе 6 сформулированы основные результаты работы.

2. Обзор литературы.

2.1. Наборы данных. В работе рассматриваются четыре открытых медицинских набора данных с различными характеристиками типа события, количества наблюдений, количества признаков, дисбаланса классов и заполненности данных.

Набор данных Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT2) [4] содержит показатели неизлечимо больных пациентов, находящихся на жизнеобеспечении. В качестве события рассматривается смерть пациента. Набор данных содержит 9105 наблюдений и 35 признаков по анамнезу, классу заболевания пациента, тяжести физиологических отклонений и сопутствующим заболеваниям. Категориальными являются 11 признаков: sex, dzgroup, dzclass, num_co, race, diabetes, dementia, ca, dnr, sfdm2, income. Пропуски содержатся в 21 признаке, максимальное количество пропусков имеет ADL (5641 пропуск). В ходе исследования были цензурированы 2904 пациента.

Набор данных WUHAN, собранный с 10 января по 18 февраля 2020 г., был представлен в работе [5]. В качестве события рассматривается время выписки пациента. Набор данных содержит 375 наблюдений и 76 признаков по анамнезу и результатам клинических исследований за время лечения. Пространство признаков формируется из минимальных, максимальных и средних показателей клинических исследований пациента. Все признаки набора данных могут содержать пропуски, максимальное количество которых имеется в показателях антитромбина и продуктах распада фибрина (173 пропуска). В ходе исследования был цензурирован 201 пациент.

Набор данных Cohort study on breast cancer patients from the Netherlands (ROTT2) [6] содержит информацию о пациентах с раком молочной железы, перенесших операцию на груди. В качестве события рассматривается рецидив рака. Набор данных содержит 2982 наблюдения и 11 признаков по анамнезу, характеристикам опухоли и стратегии лечения. Категориальными являются 6 признаков: meno, tsize, grade,

2024, **25** (3), 357–377. doi 10.26089/NumMet.v25r328

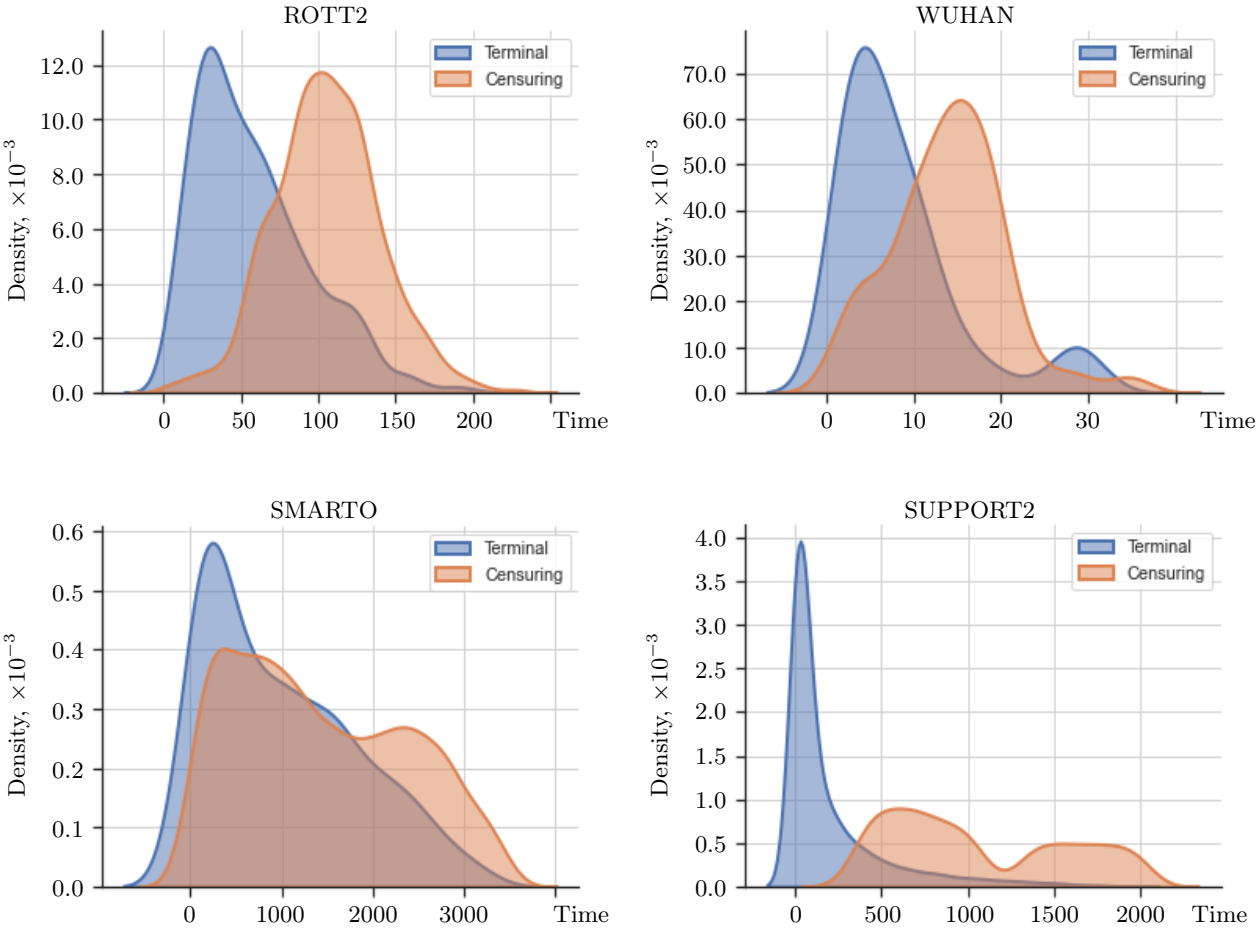


Рис. 1. Распределение времени терминальных (отмечено синим) и цензурированных (отмечено оранжевым) наблюдений для наборов данных ROTT2, WUHAN, SMARTO, SUPPORT2. Для наборов ROTT2, WUHAN наблюдается смещение распределений во времени при сохранении формы. Для наборов SMARTO и SUPPORT2 наблюдается смещение и изменение формы распределений

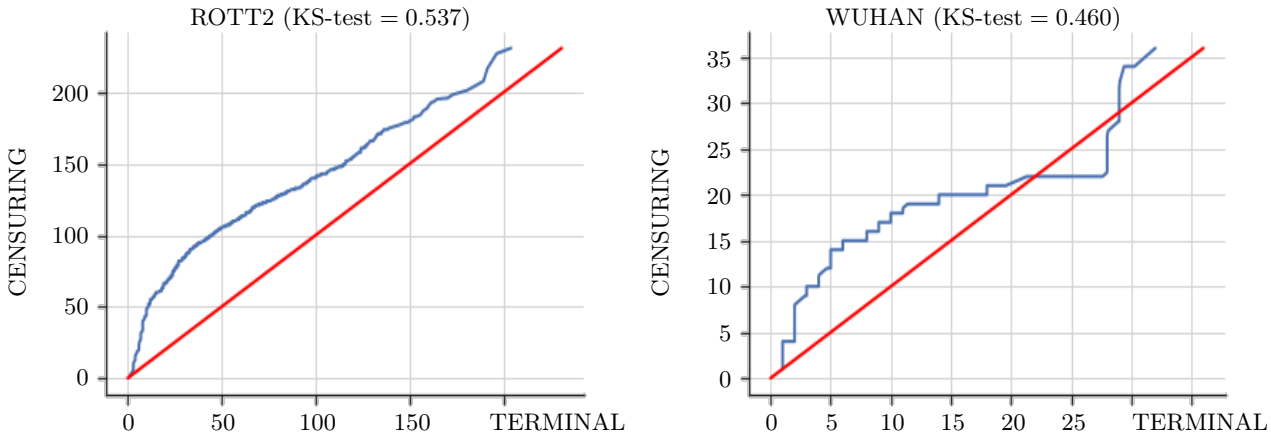
Fig. 1. Time distribution of terminal (marked in blue) and censored observations (marked in orange) for the ROTT2, WUHAN, SMARTO, SUPPORT2 datasets. For the ROTT2 and WUHAN datasets, a shift in distributions over time is observed while maintaining the shape. For the SMARTO and SUPPORT2 datasets, a shift and a change in the shape of the distributions are observed

hormon, chemo, recent. Пропусков набор данных не содержит. В ходе исследования были цензурированы 1710 пациентов.

Набор данных Second Manifestations of ARTerial Disease (SMARTO) [7] содержит сведения о пациентах, госпитализированных с клинически манифестным атеросклеротическим заболеванием сосудов или выраженными факторами риска атеросклероза. В качестве события рассматривается смерть пациента. Набор данных содержит 3873 наблюдения и 26 признаков по анамнезу, клиническим показателям и маркерам атеросклероза. Категориальными являются 9 признаков: sex, diabetes, cerebral, aaa, periph, stenosis, albumin, smoking, alcohol. Пропуски содержатся в 16 признаках, максимальное количество пропусков имеется в показателях артериального давления: diastolic by hand (1499 пропусков), systolic by hand (1498 пропусков). В ходе исследования были цензурированы 3413 пациентов.

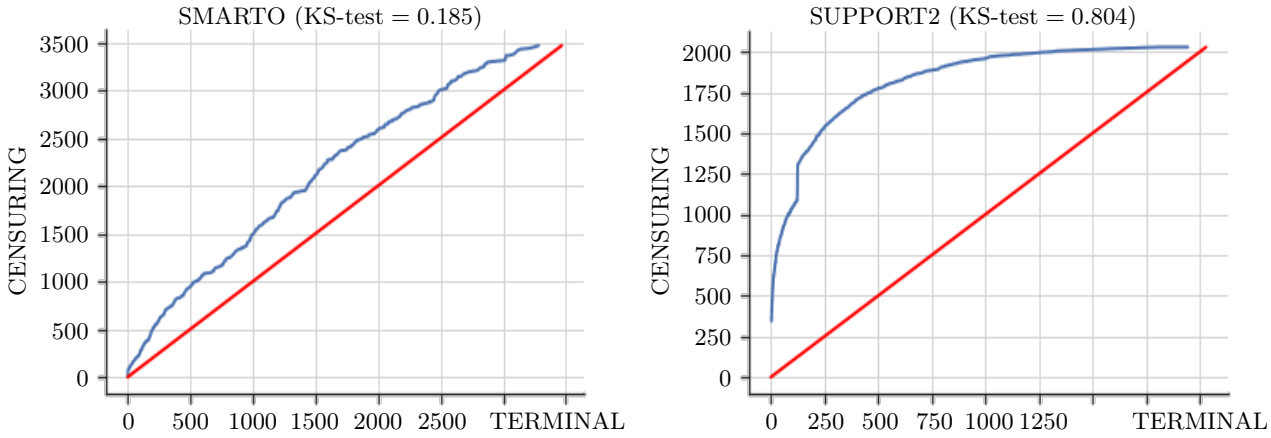
На рис. 1 представлена ядерная оценка плотности времени событий наборов данных. Ранние события наиболее важны в наборе SUPPORT2, так как набор посвящен исследованию неизлечимых пациентов на жизнеобеспечении. Дисбаланс классов событий смещен в сторону терминальных событий. Наборы данных WUHAN и ROTT2 сбалансированы относительно классов событий и наибольшую важность в них имеют ранние и средние события. Набор SMARTO имеет высокий дисбаланс цензурированных событий,

2024, **25** (3), 357–377. doi 10.26089/NumMet.v25r328



a) Сравнение распределений времени событий на наборах ROTT2, WUHAN

a) Comparison of event time distributions on ROTT2, WUHAN datasets



b) Сравнение распределений времени событий на наборах SMARTO, SUPPORT2

b) Comparison of event time distributions on SMARTO, SUPPORT2 datasets

Рис. 2. Графики квантиль-квантиль для сравнения распределения терминальных (отмечены на оси x) и цензурированных (отмечены на оси y) наблюдений во времени. Для всех наборов точки графика не лежат на прямой $y = x$ (красная линия), что говорит об информативности цензурирования в данных

Fig. 2. Quantile-quantile plots comparing the distribution of terminal (marked on the x-axis) and censored (marked on the y-axis) observations over time. For all sets, the plot points do not lie on the $y = x$ line (red line), indicating that censoring in the data is informative

которые одинаково важны за все время наблюдения. Следует отметить, что формы функций плотности распределения терминальных и цензурированных событий близки.

На рис. 2 представлено сравнение распределений терминальных и цензурированных событий. Для сравнения используется график квантиль-квантиль [8], отображающий зависимость квантилей распределения цензурированных событий (на оси y) от квантилей распределения терминальных событий (на оси x). Если два распределения схожи, то точки графика будут лежать на линии $y = x$ (отмечена красным). Однако для всех наборов данных наблюдаются значимые отклонения графика от линии $y = x$. Также для оценки максимального расстояния между эмпирическими распределениями вычислялась статистика Колмогорова–Смирнова [9]. Для всех наборов данных наблюдаются значимые различия между распределениями, что говорит об информативности индикатора цензурирования. Далее мы рассмотрим существующие модели анализа выживаемости, использующиеся для описания событийных данных и построения индивидуальных прогнозов.

2024, **25** (3), 357–377. doi 10.26089/NumMet.v25r328

2.2. Статистические модели.

§ 2.2.1. Оценка Каплана–Мейера. Наиболее распространенным методом оценки функции выживания является метод Каплана–Мейера (Kaplan–Meier, KM) [10]. Метод предполагает неинформативность флага цензурирования и строит ступенчатую оценку по всем моментам наступления события t_i. Для каждого момента времени t_i рассчитывается число оставшихся наблюдений N_i и число произошедших событий O_i.

В таком случае функцию выживания в момент t можно оценить кумулятивным произведением долей выживших по прошедшим моментам времени:

$$S(t) = \prod_{i:t_i \leqslant t} (1 - P(t_i)) = \prod_{i:t_i \leqslant t} \left(1 - \frac{O_i}{N_i}\right).$$

На основе оценки функции выживания также можно рассчитать ожидаемое время жизни $\hat{T}$: $S(\hat{T}) = 0.5$. Стоит отметить, что предположение о неинформативности цензурирования может искажать прогноз непараметрической модели. В исследовании [2] отмечается, что метод Каплана–Мейера приводит к завышению оценки функции при положительной корреляции времени события и времени цензурирования и занижению оценки функции при отрицательной корреляции.

§ 2.2.2. Оценка Нельсона–Аалена. Для оценки функции риска в анализе выживаемости используется метод Нельсона–Аалена (Nelson–Aalen, NA) [11]. Метод предполагает независимость наблюдений и цензурирования: если рассматривать два случайных объекта в момент времени $t - 1$ и один из объектов подвергается цензуре в момент времени t, а другой выживает, то оба объекта должны иметь равные шансы выжить в момент времени t.

Пользуясь терминами из раздела 2.2.1, кумулятивную функцию риска в момент t можно оценить кумулятивной суммой долей наступивших событий по прошедшим моментам времени:

$$\hat{H}(t) = \sum_{i:t_i \leqslant t} P(t_i) = \sum_{i:t_i \leqslant t} \left(\frac{O_i}{N_i}\right).$$

§ 2.2.3. Метод пропорциональных рисков Кокса. Непараметрические методы анализа выживаемости не позволяют оценить влияние признаков наблюдений на целевые переменные времени и флага цензурирования. Для преодоления данного ограничения параметрические методы предполагают теоретическую связь между признаками и целевыми переменными. На основе коэффициентов вклада переменных в расчет прогноза может быть определена значимость признаков.

Наиболее распространенным полупараметрическим методом является метод пропорциональных рисков Кокса (Cox Proportional Hazards, CoxPH) [12]. Метод основывается на допущении, что все наблюдения имеют одинаковую форму функции риска и отличаются положительным коэффициентом масштабирования, который вычисляется через скалярное произведение вектора весов и вектора признаков наблюдения:

$$h(t \mid x) = h_0(t) \exp\left(X^T \beta\right),$$

где $h_0(t)$ — базовая функция риска, X — вектор признаков, β — вектор весов линейной модели. Базовые функции риска $h_0(t)$ и кумулятивного риска $H_0(t)$ строятся на основе метода Нельсона–Аалена (раздел 2.2.2).

Обучение модели проводится через подбор линейных коэффициентов β по методу максимального правдоподобия. Прогноз индивидуальной функции выживания основан на расчете базовой функции выживания $S_0(t) = \exp\big(-H_0(t)\big)$ с последующим смещением функции с учетом коэффициента пропорциональности:

$$S(t \mid X) = \exp\left(-H_0(t) \exp(X^T \beta)\right) = S_0(t)^{\exp\left(X^T \beta\right)}.$$

Однако метод имеет несколько значимых недостатков.

- Отношение двух функций риска постоянно во времени. Следовательно, функции выживания для разных векторов признаков не пересекаются. Данное свойство может приводить к повышению значимости влияния факторов на прогноз при пересекающихся непараметрических оценках функции выживания.
- Независимость значимости признаков от времени. В клинической практике влияние факторов на риск может изменяться во времени. Например, после проведения операции пациент более подвержен риску, а после реабилитации более стабилен.

2024, **25** (3), 357–377. doi 10.26089/NumMet.v25r328

- Веса модели определяют линейную зависимость коэффициента масштабирования от исходных признаков.
- Обработка только заполненных числовых признаков. Реальные данные также могут содержать категориальные и пропущенные значения.

§ 2.2.4. Модель ускоренного времени отказа. Модель ускоренного времени отказа (Accelerated Failure Time, AFT) [13] основана на идее горизонтального масштабирования функции выживания относительно коэффициента ускорения. В частности, фактор ускорения — это константа γ, описывающая сдвиг базовой функции выживания $S_0(t)$ на основе пространства признаков наблюдения A: $S_A(t) = S_0(\gamma t)$. Коэффициент ускорения используется для сравнения времени выживания двух групп и оценки значимости признаков.

Пусть X — матрица признаков, β — вектор коэффициентов, $\sigma(\sigma > 0)$ — параметр масштабирования, а ε — случайная величина ошибки. Модель AFT предполагает неинформативность цензурирования и линейную зависимость между логарифмом времени выживания T и значениями признаков:

$$\ln T = X\beta + \sigma\varepsilon.$$

Также модель AFT учитывает теоретическое распределение времени события T, которое в современных исследованиях наиболее часто основывается на распределении Вейбулла, логистическом и логлогистическом распределениях. Распределение времени события используется для построения базовой функции выживания $S_0(t)$, функции плотности $f_0(t)$ и функции опасности $h_0(t)$, в которых оценки параметров распределения вычисляются на обучающей выборке.

Прогноз ожидаемого времени наступления события по модели AFT определяется по формуле: $T = e^{X\beta}e^{\sigma\varepsilon}$. Прогнозы функций $S(t|X)$, $f(t|X)$ и $h(t|X)$ имеют следующий вид:

$$S(t|X) = S_0(e^{-X\beta}t),\ f(t|X) = -e^{-X\beta}f_0(e^{-X\beta}t),\ h(t|X) = e^{-X\beta}h_0(e^{-X\beta}t).$$

Таким образом, метод AFT имеет ряд существенных недостатков.

- Строгое предположение о теоретическом распределении времени.
- Предположение о масштабируемости функций выживания по времени ограничивает возможность пересечения прогнозов функций выживания.
- Веса модели определяют линейную комбинацию исходных признаков.
- Обработка только заполненных числовых признаков.

2.3. Древовидные модели машинного обучения. Древовидные подходы позволяют преодолеть строгие теоретические предположения статистических моделей: пропорциональность риска, линейность зависимостей, неинформативность индикатора цензурирования. Эти подходы основаны на идее рекурсивного разбиения признакового пространства на области с близкими значениями целевой переменной.

Критерии разбиения выборок с цензурированием разделяются на два типа [14]. Первый тип основан на идее минимизации ошибки описания данных в дочерних выборках. Для каждой пары ветвей разбиения строится непараметрическая оценка функции выживания и выбирается лучшее разбиение с минимальным отрицательным правдоподобием или расстоянием Васерштейна [15]. Критерии второго типа основаны на максимизации расстояния между выборками путем сравнения дочерних функций выживания или функций риска. Наиболее популярный критерий основан на вычислении статистики log-rank [16].

§ 2.3.1. Критерий log-rank. Для измерения различий между функциями выживания двух групп наибольшее распространение получил критерий log-rank [16]. Большее значение статистики log-rank определяет большее различие между двумя выборками. Нулевая гипотеза критерия H_0 предполагает, что функции риска двух выборок совпадают: $h_1(t) = h_2(t)$.

Пусть даны две группы с n_1 и n_2 наблюдениями. Определим упорядоченный набор времени наступления событий: $\tau_1 < \tau_2 < \ldots < \tau_K$. Пусть $N_{1,j}$ и $N_{2,j}$ — количество наблюдений на момент τ_j, а $O_{1,j}$ и $O_{2,j}$ — количество событий в момент τ_j. Тогда общее число наблюдений и событий на момент τ_j: $N_j = N_{1,j} + N_{2,j}$ и $O_j = O_{1,j} + O_{2,j}$ соответственно. Определим ожидаемое число событий на момент τ_j как $E_{i,j} = N_{i,j}O_j/N_j$. На основе имеющихся данных можно рассчитать статистику log-rank:

$$\mathrm{LR} = \sum_{j=1}^{K} w_j\,(O_{1,j} - E_{1,j}) \Bigg/ \sqrt{\sum_{j=1}^{K} w_j^2 E_{1,j} \left(\frac{N_j - O_j}{N_j}\right)\left(\frac{N_j - N_{1,j}}{N_j - 1}\right)}, \tag{1}$$

2024, **25** (3), 357–377. doi 10.26089/NumMet.v25r328

где $w_j = 1$. Тест log-rank обладает оптимальной мощностью для обнаружения различий в выборках, в которых функции риска пропорциональны друг другу.

Однако в исследованиях [16, 17] высказывается предположение о плохой чувствительности log-rank к реальным данным с доминирующим числом ранних событий. Критерий log-rank основывается на предположении, что индикатор цензурирования неинформативен, вероятности выживания одинаковы для ранних и поздних событий.

Существует несколько модификаций [16] весов критерия w_j, которые повышают значимость ранних событий: wilcoxon $w_j = N_j$, peto-peto $w_j = \hat{S}(\tau_j)$ (здесь $\hat{S}(\tau_j)$ — оценка функции выживания по методу Каплана-Мейера (раздел 2.2.1)), tarone-ware $w_j = \sqrt{N_j}$.

§ 2.3.2. Дерево выживания. Классический алгоритм дерева решения [18] основан на идее рекурсивного разделения выборки на группы с разной выживаемостью. Используя заранее определенный критерий, корневой узел (содержит все данные) разделяется на два дочерних узла. Процесс повторяется рекурсивным образом для каждого из дочерних узлов.

Метод построения дерева выживания [19] является расширением классического метода с возможностью прогнозирования величин анализа выживаемости. При разбиении узла по каждому признаку из множества X рассматриваются всевозможные промежуточные значения. Для каждого значения строятся две ветви разбиения и вычисляется значение статистики log-rank. Лучшее разбиение выборки выбирается по максимальному значению статистики. Прогноз для наблюдения с вектором признаков x вычисляется на основе данных, которые находятся в том же листе (конечном узле), что и x. Для прогноза функции выживания используется оценка Каплана–Мейера, а функции риска — оценка Нельсона–Аалена.

Также для контроля вычислительной сложности модели дерево выживания поддерживает аппарат ограничения роста (pre-pruning) путем задания структуры дерева с помощью гиперпараметров [20]: максимальное количество листов, максимальная глубина.

Преимуществом метода является сильная интерпретация. Каждому полученному листу сопоставляется набор правил при проходе от корня к листу. Для оценки корректности прогноза эксперту достаточно проанализировать последовательность правил разделения данных.

Однако существующая реализация имеет значимые недостатки. Во-первых, построение дерева возможно только на заполненных данных. Во-вторых, дерево выживания наследует недостатки критерия log-rank. Наконец, для построения точного дерева необходимо достаточное количество данных.

2.4. Метрики качества. Существует несколько способов оценки качества прогнозирования величин анализа выживаемости. Каждая метрика качества позволяет оценивать одну из прогнозируемых величин: ожидаемое время события T, функцию выживания $S(t)$, функцию риска $h(t)$.

§ 2.4.1. Concordance index. Concordance index (CI) [21] измеряет долю верно упорядоченных пар относительно времени наступления события. Лучшее значение метрики равно 1 (при полностью верном упорядочивании), худшее значение равно 0 (все пары упорядочены неверно), а значение 0.5 отражает случайность отклика модели. Пусть T_k — истинное время наступления события, а η_k — прогноз времени, тогда

$$\mathrm{CI} = \frac{\sum\limits_{i,j} 1_{T_j<T_i} \cdot 1_{\eta_j<\eta_i}}{\sum\limits_{i,j} 1_{T_j<T_i}}. \tag{2}$$

Однако метрика основана на точечном прогнозе времени события без учета свойств функций выживания и риска. В то же время значение CI не изменяется при сдвиге функций выживания, хотя прогнозируемое время будет сильно искажено по сравнению с истинным.

§ 2.4.2. Integrated AUC. Альтернативной метрикой ранжирования является integrated AUC. В работе [22] представлен метод распространения вычисления ROC-кривой и площади под кривой (AUC) на многоклассовые или временны́е случаи. Для всех моментов t временно́й шкалы определяются множества наступивших и ненаступивших событий. Метрика $\widehat{\mathrm{AUC}}(t)$ измеряет качество ранжирования пар наблюдений из каждого множества на основе кумулятивного риска в момент t:

$$\widehat{\mathrm{AUC}}(t) = \frac{\sum\limits_i \sum\limits_j I\left(T_j > t \geqslant T_i\right) I\left(\hat{h}_j(t) \leqslant \hat{h}_i(t)\right) \delta_i w_i}{\left(\sum\limits_j I(T_j > t)\right)\left(\sum\limits_i I(T_i \leqslant t) \delta_i w_i\right)},$$

2024, **25** (3), 357–377. doi 10.26089/NumMet.v25r328

где $\hat{h}_i(t) = \hat{h}(t, x_i)$ — оценка кумулятивного риска наблюдения x_i в момент времени t, $w_i = P(C > t_i)$ — обратная вероятность цензурирования наблюдения x_i по оценке Каплана–Мейера, δ_i — флаг цензурирования наблюдения x_i.

Для агрегации оценок $\widehat{\mathrm{AUC}}(t)$ во времени используется следующая метрика:

$$\mathrm{IAUC} = \frac{1}{\int w(t)dt} \int \widehat{\mathrm{AUC}}(t) w(t) dt. \tag{3}$$

В работах [23, 24] рассматривается взвешивание на основе функции плотности $w(t) = \hat{f}(t)$. В исследовании чувствительности метрик [25] отмечается наличие повышенного вклада поздних моментов времени в метрику IAUC. Для преодоления предвзятости предлагается использовать $\mathrm{IAUC}(w(t) = 1)$ с равным вкладом $w_i = 1$ частных наблюдений и интегрированием с весовой схемой $w(t) = 1$.

§ 2.4.3. Integrated Brier Score. Метрика Integrated Brier Score [26, 27] основана на расчете квадратичного отклонения прогнозируемой функции выживания от истинной. Для события с временем T истинная функция выживания равна 1 до момента наступления события ($T > t$) и 0 после ($T \leqslant t$).

Для метрики отмечается наличие повышенного вклада поздних событий независимо от распределения событий [25]. Для преодоления смещения предлагается модификация $\mathrm{IBS}_{\mathrm{RM}}$ с равным весом наблюдений и контролируемым усреднением наблюдаемых событий $N(t)$ в момент времени t. Пусть $S(t \mid X_i)$ — прогноз функции выживания в момент t для наблюдения X_i с временем наступления события T_i, тогда:

$$\mathrm{BS}_{\mathrm{RM}}(t) = \frac{1}{N(t)} \sum_i \begin{cases} (0 - S(t \mid X_i))^2, & T_i \leqslant t,\ \delta_i = 1, \\ (1 - S(t \mid X_i))^2, & T_i > t, \\ 0, & T_i = t,\ \delta_i = 0, \end{cases}$$

$$\mathrm{IBS}_{\mathrm{RM}} = \frac{1}{t_{\max}} \int_0^{t_{\max}} \mathrm{BS}_{\mathrm{RM}}(t) dt. \tag{4}$$

§ 2.4.4. Area under the Precision-Recall Curve. Метрика Survival-AUPRC (AUPRC) [28] основана на идее измерения качества пороговости функции выживания в различных окрестностях момента времени наступления события. В случае терминальных событий оценка качества сводится к усреднению качества прогнозирования при разных промежутках $[T_i \cdot \varphi, T_i/\varphi]$, где $\varphi \in [0, 1]$. В случае цензурированных событий рассматривается только оценка функции выживания до момента наступления события:

$$\mathrm{AUPRC}_{\delta_i = 1}(\hat{S}, T_i) = \int_0^1 \hat{S}(T_i \cdot \varphi) - \delta_i \cdot \hat{S}(T_i/\varphi) d\varphi.$$

Лучшее значение достигается, если функция выживания представляет собой пороговую функцию, равную 1 до наступления события и 0 после. Наименьшее значение достигается, если функция выживания является константой (любая константная функция для терминальных событий и константный 0 для цензурированных наблюдений). Для оценки качества по выборке рассматривается среднее значение:

$$\mathrm{AUPRC} = \frac{1}{N} \sum_i \mathrm{AUPRC}_{\delta_i}(\hat{S}(X_i), T_i). \tag{5}$$

3. Чувствительность древовидных подходов. Наличие информативности цензурирования приводит к различиям между распределениями терминальных и цензурированных наблюдений. При наличии двух распределений с различной формой и местоположением возникают выборки с мультимодальным распределением времени. В частности, в рассмотренных ранее медицинских данных наблюдаются случаи мультимодального распределения времени (рис. 1).

При разделении исходной выборки возникает проблема появления мультимодальных распределений в подмножествах. На рис. 3 представлен пример построения дерева решений на наборе данных SMARTO. Для каждого узла дерева представлена ядерная оценка плотности времени событий, а также сводная информация о выборке: размер узла (size), доля терминальных событий (event/size), среднее

2024, **25** (3), 357–377. doi 10.26089/NumMet.v25r328

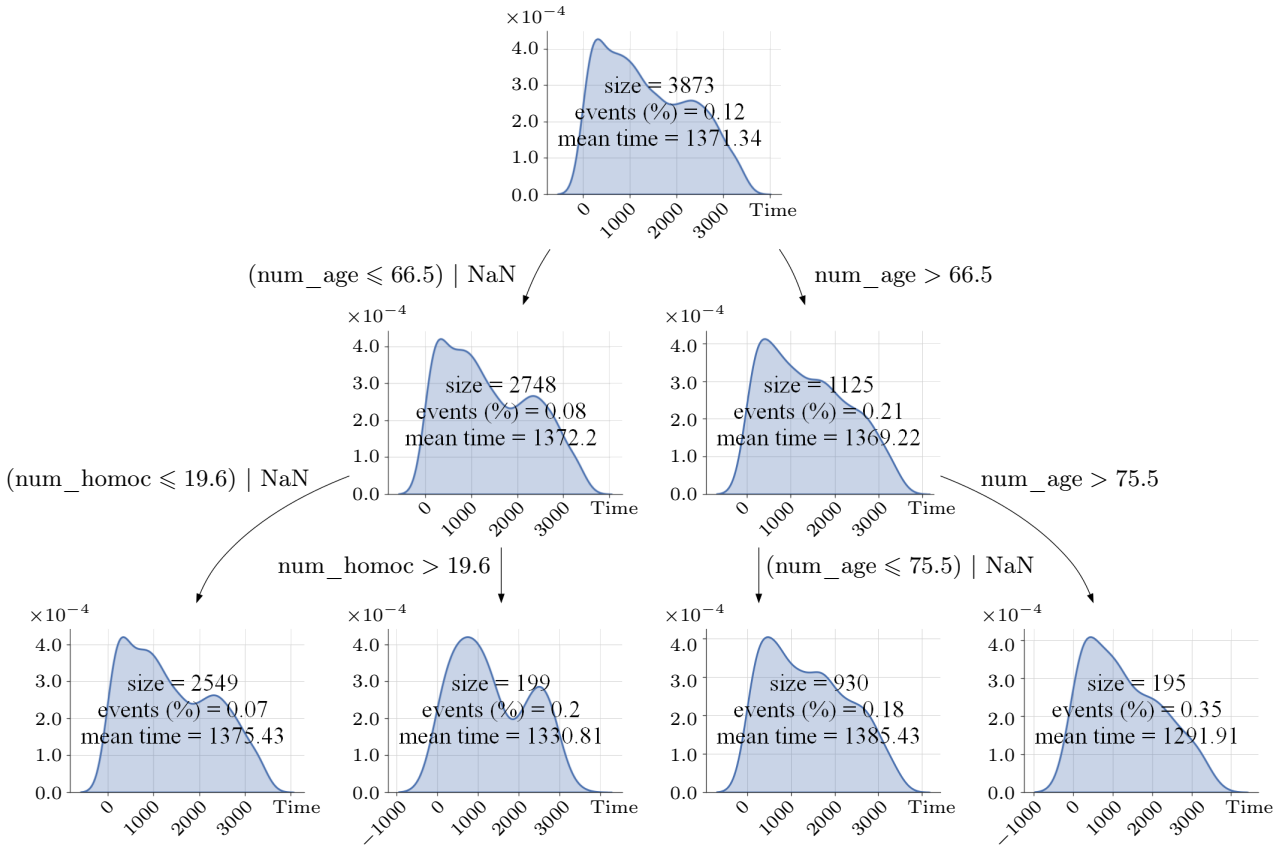


Рис. 3. Пример дерева выживания, построенного на наборе данных SMARTO с гиперпараметрами: глубина равна 2, минимальный размер листа равен 0.01 и критерий разбиения log-rank. В построенном дереве использовались признаки: num_age (возраст) и hum_homoc (гомоцистеин, μ моль/л)

Fig. 3. An example of a survival tree built on the SMARTO dataset with hyperparameters: depth equals 2, minimum leaf size equals 0.01, and log-rank split criterion. The constructed tree used features: num_age (age) and hum_homoc (homocysteine, μ mol/L)

время наступления события (time of event). В результате разбиения исходной выборки листовые узлы имеют унимодальное и мультимодальное распределение времени событий. В частности, лист с правилом $\big((\text{num\_age} \leqslant 66.5) \mid \text{NaN}\big)$ & $(\text{num\_homoc} > 19.6)$ имеет мультимодальное распределение.

Одной из причин появления мультимодальных подвыборок являются признаки, связанные со схемой лечения. Например, летальный исход для некоторых пациентов наступает на начальном этапе после операции, а для других — намного позже этапа реабилитации. Признаки диагностики заболеваний также неявно указывают на схему лечения. Попадание пациента в зону риска (например, на основе результатов клинических анализов или анамнеза) требует врачебного вмешательства и приводит к стратификации пациентов относительно успешности схемы лечения. Важно отметить, что критерий разбиения может повышать значимость разбиений с мультимодальным распределением времени.

3.1. Чувствительность критерия log-rank. Классический критерий log-rank предполагает независимость и случайность обеих выборок, а также неинформативность индикатора цензурирования. При работе с малыми выборками критерий также ограничен временно́й шкалой и не чувствителен к поздним событиям, наблюдаемым после исчерпания одной из выборок. Для демонстрации недостатков критерия мы вновь рассмотрим log-rank статистику (1). Далее все обозначения согласованы с параграфом 2.3.1.

Пусть максимальное время события в первой группе равно τ_{K1}, а во второй τ_{K2}, причем $K1 \leqslant K2$ (в противном случае поменяем группы местами). Тогда для $\tau_k \in (\tau_{K1}, \tau_{K2}]$ первая группа не содержит событий и $O_{1,k} = 0$, $N_{1,k} = 0$. Следовательно, ожидаемое количество событий $E_{i,k} = O_{i,k}$ и вклад отклонений в числитель и знаменатель равен 0. Таким образом, после наступления всех событий одной из групп вклад поздних событий другой группы не учитывается.

2024, **25** (3), 357–377. doi 10.26089/NumMet.v25r328

Данный эффект влияет на появление подвыборок с мультимодальным распределением, поскольку критерий использует ограниченную информацию о наблюдаемых данных. Рассмотрим следующий пример. Пусть дана выборка терминальных событий с временем от 0 до 150 с шагом 10. Пусть после разбиения выборки в первую группу вошли события с временем $[0, 10, 20, 30, 80, 90, 100]$, а во вторую — с временем $[40, 50, 60, 70, 110, 120, 130, 140, 150]$. Таким образом, исходная выборка имеет равномерное распределение времени. Группы разбиения имеют мультимодальное распределение.

При расчете log-rank статистики количество наблюдений $N_{1,j}$ и $N_{2,j}$ и ожидаемое количество событий $E_{1,j}$ имеют следующие значения (значения округлены для удобства восприятия):

$$\boldsymbol{N}_{1,j} = [7, 6, 5, 4, 3, 3, 3, 3, 3, 2, 1, \mathbf{0,0,0,0,0}],$$

$$\boldsymbol{N}_{2,j} = [9, 9, 9, 9, 9, 8, 7, 6, 5, 5, 5, \mathbf{5,4,3,2,1}],$$

$$\boldsymbol{E}_{1,j} = [0.44, 0.4, 0.36, 0.31, 0.25, 0.27, 0.3, 0.33, 0.38, 0.29, 0.16, \mathbf{0,0,0,0,0}].$$

Жирным выделены моменты времени с поздними событиями второй группы, которые имеют нулевой вклад из-за наступления всех событий первой группы ($N_{2,j} = N_j, E_{2,j} = O_{1,j}$). С точки зрения критерия log-rank разбиение является значимым (уровень значимости 0.05): статистика критерия равна 5.313, а p-value — 0.0211.

Однако значимость разбиений с мультимодальными группами не устойчива по критерию log-rank. Добавим к каждой группе по одному цензурированному наблюдению в момент времени $\tau = 160$. Обновленные значения $N_{1,j}$ и $N_{2,j}$ имеют следующий вид (жирным выделены моменты времени, которые ранее имели нулевой вклад):

$$\boldsymbol{N}_{1,j} = [8, 7, 6, 5, 4, 4, 4, 4, 4, 3, 2, \mathbf{1,\ 1,\ 1,\ 1,\ 1}, 0],$$

$$\boldsymbol{N}_{2,j} = [10, 10, 10, 10, 10, 9, 8, 7, 6, 6, 6, \mathbf{6,\ 5,\ 4,\ 3,\ 2}, 1].$$

Данное разбиение является незначимым: статистика критерия равна 1.298, а p-value — 0.2545. Разительное отличие статистик связано с тем, что в обновленных данных поздние события имеют ненулевой вклад при расчете статистики. Таким образом, по критерию log-rank существуют значимые разбиения, которые приводят к появлению двух групп с мультимодальным распределением времени. Для использования полной информации о наблюдаемых данных необходимо учитывать вклад поздних событий даже при исчерпании одной из выборок.

3.2. Модификация весов log-rank. Критерий log-rank применяется для обнаружения различий в выборках с пропорциональными функциями риска. Как отмечено выше, для повышения значимости определенного интервала времени событий настраиваются веса w_j критерия log-rank (1).

Ранние события имеют высокую значимость в критериях wilcoxon ($w_j = N_j$), tarone-ware ($w_j = \sqrt{N_j}$), peto-peto ($w_j = \hat{S}(\tau_j)$). Критерий fleming-harington [29] повышает значимость определенного интервала времени на основе семейства статистик $\{G^{\rho,\gamma} : \rho \geqslant 0, \gamma \geqslant 0\}$ с весовой схемой $w_j = \hat{S}(\tau_j)^\rho \big(1 - \hat{S}(\tau_j)\big)^\gamma$. Критерий $G^{0,0}$ равен критерию log-rank (чувствительность к пропорциональным рискам), $G^{1,0}$ равен peto-peto критерию (чувствительность к ранним событиям), а $G^{0,1}$ чувствителен к поздним событиям [30].

Однако повышение значимости поздних событий не всегда приводит к повышению чувствительности критерия. В разделе 3.1 представлен случай обнуления вклада поздних событий при исчерпании событий одной из выборок. В таком случае веса w_j не влияют на величину вклада поздних событий. Следовательно, настройка весов критерия log-rank не позволяет избежать появления выборок с мультимодальным распределением времени.

4. Предлагаемый подход. В разделе 2 мы рассмотрели случаи и последствия появления выборок с мультимодальным распределением. Низкое качество прогнозирования и сложность оценки ожидаемого времени ограничивают применимость существующих моделей анализа выживаемости. В данном разделе представлена модификация метода построения деревьев выживания, которая преодолевает недостатки мультимодального распределения времени событий.

Модификация основана на новом подходе к регуляризации критериев разбиения. Регуляризованные критерии учитывают априорную информацию о распределении наблюдений и событий при поиске наилучшего разбиения. Предлагаемый подход встраивается в метод построения деревьев выживания на примере взвешенных критериев log-rank.

2024, **25** (3), 357–377. doi 10.26089/NumMet.v25r328

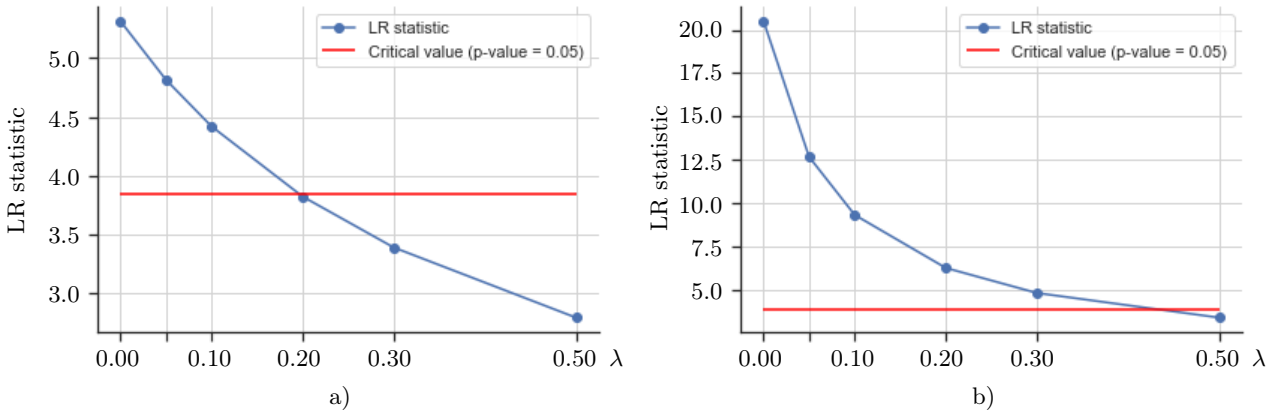


Рис. 4. Оценка изменения значимости разбиения с увеличением коэффициента регуляризации: a) разбиение с мультимодальным распределением групп; b) разбиение с унимодальным распределением

Fig. 4. Evaluation of the change in the significance of partitioning with an increase in the regularization coefficient: a) partitioning with a multimodal distribution of groups; b) partitioning with a unimodal distribution

4.1. Регуляризация критерия log-rank. Как говорилось ранее, современные способы борьбы с переобучением деревьев решений позволяют ограничить сложность дерева на этапе построения модели или обрезки избыточных листов. Однако эти способы не используют информацию о распределении событий в выборке и не позволяют преодолеть проблему обнуления вклада поздних событий (см. раздел 3.1).

В данном разделе предлагается метод регуляризации статистики log-rank для повышения чувствительности критерия к распределению событий. Метод позволяет учитывать информацию об априорном распределении событий при сравнении выборок с цензурированием. Формально подход основан на добавлении информации об априорном распределении оставшихся и наступивших событий для всех моментов времени. Априорная информация добавляется с коэффициентом регуляризации к наблюдаемым распределениям в выборках и определяет ненулевой вклад всех интервалов времени.

Обозначим через N_j^A априорное количество оставшихся наблюдений к моменту τ_j, O_j^A — априорное количество событий в момент τ_j. Введем дополнительный параметр регуляризации λ. Аналогично формуле критерия log-rank (раздел 3.1) рассмотрим две группы с количеством оставшихся наблюдений $N_{1,j}$, $N_{2,j}$, количеством событий $O_{1,j}$, $O_{2,j}$. Введем обновленные значения $\hat{N}_{1,j}, \hat{N}_{2,j}$, $\hat{O}_{1,j}, \hat{O}_{2,j}$ согласно следующим формулам:

$$\hat{N}_{i,j} \leftarrow N_{i,j} + \frac{\lambda}{2} \cdot N_j^A,\ \hat{O}_{i,j} \leftarrow O_{i,j} + \frac{\lambda}{2} \cdot O_j^A,$$

$$\hat{N}_j \leftarrow \hat{N}_{1,j} + \hat{N}_{2,j},\ \hat{O}_j \leftarrow \hat{O}_{1,j} + \hat{O}_{2,j}.$$

Определим математическое ожидание и дисперсию числа событий на момент τ_j как $\hat{E}_{i,j} = \dfrac{\hat{N}_{i,j}\hat{O}_j}{\hat{N}_j}$ и $\hat{V}_{i,j} = \hat{E}_{i,j}\left(\dfrac{\hat{N}_j - \hat{O}_j}{\hat{N}_j}\right)\left(\dfrac{\hat{N}_j - \hat{N}_{i,j}}{\hat{N}_j - 1}\right)$. Для оценки значимости разбиения с регуляризацией используем следующий вид статистики log-rank, основанный на нулевой гипотезе $\hat{h}_1(t) = \hat{h}_2(t)$:

$$\mathrm{LR} = \frac{\sum_{j=1}^{K} w_j \left(\hat{O}_{1,j} - \hat{E}_{1,j}\right)}{\sqrt{\sum_{j=1}^{K} w_j^2 \hat{V}_{1,j}}}. \tag{6}$$

Приведенная выше модификация критерия log-rank позволяет учитывать априорную форму распределения при $\lambda > 0$. На рис. 4 представлен пример изменения значимости двух разбиений. На левом графике представлено разбиение с мультимодальным распределением групп из примера раздела 3.1. На правом графике представлено разбиение с унимодальным распределением групп (первая группа содержит события со временем $[0, 10, 20, 30]$, а вторая — $[40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150]$). Красная

2024, **25** (3), 357–377. doi 10.26089/NumMet.v25r328

линия определяет уровень значимости статистики log-rank. Видно, что значимость разбиений убывает, однако разбиение с унимодальностью групп имеет бóльшую устойчивость к эффекту регуляризации.

Метод регуляризации критерия log-rank позволяет учесть все интервалы времени на этапе поиска лучшего разбиения. Недостатком подхода является необходимость подбора гиперпараметра λ. При высоких значениях параметра значимость разбиения уменьшается и группы становятся неразличимы.

4.2. Дерево выживания с взвешенными регуляризованными критериями. Предложенный подход к регуляризации критерия log-rank встроен в разработанный метод дерева выживания со взвешенными критериями [31]. Метод работает с категориальными и пропущенными значениями и строит модели с повышенной чувствительностью к особенностям данных. Сочетание регуляризации со взвешенным критерием log-rank позволяет учитывать различия между функциями риска на всей временно́й шкале и одновременно определять значимость определенных событий.

Для каждого доступного признака в узле используется следующий алгоритм поиска лучшего разбиения. Для непрерывных признаков определяются промежуточные точки $a_1, a_2, \ldots, a_k$ так, что $v_1 < a_1 < v_2 < a_2 < \ldots < a_{n-1} < v_n$, где $a_i = \dfrac{v_i + v_{i-1}}{2}$, а v_i — упорядоченные уникальные значения. Для каждой точки разделения a_i определяются две ветви разбиения: левая при $v \leqslant a_i$, правая при $v > a_i$. В работе предлагается ограничить количество точек разбиений k (по умолчанию 100). Если $n < k$, то в качестве точек разбиения берутся промежуточные точки, иначе значения признака дискретизируются и в качестве точек разбиения a_i принимаются $\dfrac{i}{k}$ процентили значений признака. Таким образом, при росте уникальных значений признака мощность множества промежуточных точек постоянна.

При обработке категориальных признаков рассматриваются все непересекающиеся множества l, r уникальных значений признака. Наблюдения левой и правой ветви определяются на основе правил $v \in \{l\}$ и $v \in \{r\}$ соответственно. Для контроля сложности мы используем преобразование Weight Of Evidence (WOE) [32] категориальных значений на непрерывную шкалу. Метод WOE сопоставляет каждой категории B признака значимость $w_B = \ln \dfrac{P(B \mid E)}{P(B \mid \bar{E})}$ в модели бинарной классификации. Здесь $P(B \mid E)$ определяет вероятность наблюдения категории B при условии наступления события и $P(B \mid \bar{E})$ — вероятность наблюдения категории B при условии ненаступления события.

Для всех пар ветвей вычисляется значение статистики взвешенного критерия log-rank. Пропущенные значения (при наличии) добавляются по очереди в каждую из выборок и помещаются согласно разбиению с наименьшим p-value. Применяя поправку Бонферрони [33], лучшее разбиение выбирается по минимальному p-value среди доступных признаков в узле. Поправка снижает значимость широко распространенных признаков и отдает предпочтение редким разбиениям. Алгоритм повторяется рекурсивно для каждого узла дерева.

Каждому листу дерева выживания сопоставляется оценка функции выживания (модель Каплана–Мейера), функции риска (модель Нельсона–Аалена), ожидаемая вероятность и время наступления события. В работе рассмотрены 4 весовые схемы: log-rank, wilcoxon, peto-peto, tarone-ware. По результатам экспериментального исследования, взвешенные критерии позволяют повысить гибкость модели к данным с различным распределением времени событий. Однако, из-за отсутствия универсального критерия, подбор весовой схемы необходимо проводить по исходным данным.

Для борьбы с переобучением предложенный подход также поддерживает алгоритм обрезки (post-prunning) [20]. Обрезка предполагает уменьшение размера деревьев решений за счет удаления некритичных и избыточных участков дерева. Также уменьшением размера дерева решается проблема большого размера итоговой модели.

При алгоритме обрезки входная выборка разделяется в пропорции 80% к 20% на обучающую и валидационную выборки соответственно. После построения дерева решений на обучающей выборке применяется итерационный алгоритм обрезки на валидационной выборке: на первом шаге алгоритма выделяют все поддеревья без одного листового разбиения, на втором рассчитывают метрики качества для поддеревьев, выделенных на предыдущем шаге, на третьем выбирают поддеревья с наилучшим качеством и затем возвращаются к первому шагу. Наконец, среди всех лучших поддеревьев каждой итерации выбирается модель с высшим качеством.

Также предложенный подход позволяет использовать следующие гиперпараметры для контроля роста дерева (pre-pruning): максимальная глубина дерева, максимальное количество признаков при поиске лучшего разбиения, максимальное количество наблюдений в узле.

2024, **25** (3), 357–377. doi 10.26089/NumMet.v25r328

5. Экспериментальное исследование. В данном разделе представлено экспериментальное исследование качества моделей анализа выживаемости на четырех медицинских наборах данных (раздел 2.1). Первой целью экспериментального исследования является оценка влияния коэффициента регуляризации на качество прогнозирования предложенной модели. Также мы сравним предложенный алгоритм дерева выживания с существующей реализацией дерева выживания Survival Tree (раздел 2.3.2), методом пропорциональных рисков Кокса Cox Proportional Hazards (раздел 2.2.3), моделью ускоренного времени отказа Accelerated Failure Time (раздел 2.2.4) и непараметрической моделью Kaplan–Meier (раздел 2.2.1).

5.1. Постановка эксперимента. Алгоритм эксперимента разделяется на три этапа. Первоначально проводится предобработка набора данных, формирование признакового пространства и целевых переменных (время до наступления события, индикатор цензурирования). На первом этапе исходные данные разбиваются на тренировочную и тестовую выборки (66% и 33% соответственно) со стратификацией по индикатору цензурирования и времени.

На втором этапе проводится пятикратная кросс-валидация [34] по сетке гиперпараметров на тренировочной выборке. При кросс-валидации исходная выборка разделяется на пять непересекающихся частей, четыре из которых используются для обучения модели, а одна часть используется для тестирования модели и вычисления метрик качества. Всего проводится пять итераций обучения/тестирования модели, причем каждая часть единожды используется для оценки качества на тестовой выборке. Качество модели по кросс-валидации оценивается как среднее значение метрики по всем итерациям. Данный этап включает выбор лучших гиперпараметров для каждой модели.

На третьем этапе исходные данные 20-кратно разбиваются на тренировочные и тестовые данные (66% и 33% соответственно). Лучшие модели (выбранные в ходе кросс-валидации) обучаются на тренировочных данных и применяются к тестовым. Итоговое качество модели вычисляется как среднее качество для 20 тестовых выборок.

Реализации моделей CoxPH Survival Analysis, Survival Tree были взяты из открытой библиотеки *scikit-survival* [35], а реализации моделей Kaplan–Meier и Accelerated Failure Time были взяты из библиотеки *lifelines* [36].

Реализация модели TREE представлена в разделе 4.2 данной работы. Модель поддерживает метод регуляризации, описанный в разделе 4.1. Полный список гиперпараметров используемых моделей представлен в табл. 1. Предложенная модель TREE встроена в библиотеку анализа выживаемости *survivors* [37] с открытым исходным кодом.

5.2. Результаты. В табл. 2–5 представлены результаты экспериментального сравнения классических моделей (Kaplan–Meier, Cox Proportional Hazards, Accelerated Failure Time, Survival Tree) и предложенного дерева выживания. Модель TREE поддерживает предложенный подход к регуляризации для

Таблица 1. Гиперпараметры моделей прогнозирования

Table 1. Hyperparameters of forecasting models

Название модели Model	Гиперпараметры Hyperparameters	Сетка значений Values grid
Cox Proportional Hazards (CoxPH)	regularization penalty ties	0.1, 0.01, 0.001 breslow, efron
Accelerated Failure Time (AFT)	penalizer l1 ratio distribution	0, 0.01, 0.1, 0.5, 1.0 10^x for x from -3 to 3 Weibull, LogNormal, LogLogistic
Survival Tree (ST)	split strategy max depth min sample leaf max features	best, random from 10 to 30 step 5 from 1 to 20 step 1 sqrt, log2, None
TREE	max depth min sample leaf signif lambda criterion	from 10 to 30 step 5 0.05, 0.01, 0.001 0.05, 0.1, 1.0 0.0, 0.01, 0.1, 0.5, 0.9 peto-peto, tarone-ware, wilcoxon, log-rank

2024, **25** (3), 357–377. doi 10.26089/NumMet.v25r328

поиска лучшего разбиения (значение коэффициента λ указано в названии модели). Оценка качества проводилась по метрикам: CI $\uparrow$ (2), IBS_{RM} $\downarrow$ (4), IAUC $\uparrow$ (3), AUPRC $\uparrow$ (5). В ячейках таблицы отображены средние значения метрики на 20 тестовых выборках для каждого метода. В столбце "Название модели" жирным шрифтом выделены предложенные методы, а в остальных столбцах жирным шрифтом выделены лучшие значения для каждой метрики.

Таблица 2. Сравнение качества прогнозирования на наборе ROTT2

Table 2. Comparison of forecast quality on the ROTT2 dataset

Название модели Model	CI	IBS_{RM}	IAUC	AUPRC
CoxPH	0.5960	0.4756	0.7105	0.0000
KM	0.5000	0.1791	0.4962	0.5952
LogLogisticAFT	0.6060	0.1531	0.7282	0.6291
LogNormalAFT	0.5964	0.1573	0.7106	0.6252
WeibullAFT	0.6022	0.1525	0.7187	0.6308
ST	0.5554	0.1800	0.6185	0.6075
TREE($\lambda = 0$)	0.6188	0.1282	0.7354	0.7225
TREE($\lambda = 0.01$)	0.6197	0.1281	0.7359	**0.7228**
TREE($\lambda = 0.1$)	0.6198	0.1282	0.7325	**0.7230**
TREE($\lambda = 0.5$)	**0.6234**	**0.1278**	**0.7379**	0.7225
TREE($\lambda = 0.9$)	**0.6240**	**0.1279**	**0.7401**	0.7223

Таблица 3. Сравнение качества прогнозирования на наборе WUHAN

Table 3. Comparison of forecasting quality on the WUHAN dataset

Название модели Model	CI	IBS_{RM}	IAUC	AUPRC
CoxPH	0.7088	0.1256	0.8204	0.7276
KM	0.5000	0.1913	0.4871	0.5515
LogLogisticAFT	0.6658	0.1885	0.4871	0.5099
LogNormalAFT	0.6530	0.2401	0.7347	0.6313
WeibullAFT	0.5000	0.1767	0.4871	0.5567
ST	0.6363	0.1440	0.7508	0.6514
TREE($\lambda = 0$)	0.6926	0.1010	0.8217	0.7570
TREE($\lambda = 0.01$)	0.6988	0.1003	0.8225	0.7589
TREE($\lambda = 0.1$)	**0.7108**	**0.0985**	0.8324	**0.7617**
TREE($\lambda = 0.5$)	0.7078	0.0996	**0.8349**	**0.7601**
TREE($\lambda = 0.9$)	**0.7130**	**0.0970**	**0.8328**	0.7577

Для наборов данных ROTT2 (табл. 2) и SMARTO (табл. 4) качество CI, IBS_{RM}, IAUC улучшается с увеличением λ, а качество AUPRC не зависит от степени регуляризации. Для наборов данных WUHAN (табл. 3) и SUPPORT2 (табл. 5) лучшее качество по всем метрикам достигается при ненулевом λ. По совокупности метрик лучшее качество для наборов ROTT2 и WUHAN достигается с помощью модели TREE с $\lambda = 0.9$, для набора SMARTO — TREE с $\lambda = 0.5$, а для набора SUPPORT2 — TREE с $\lambda = 0.01$.

Предложенная модель TREE превзошла существующие методы дерева выживания и Каплана–Мейера на всех наборах данных. Кроме того, модель TREE превзошла модели CoxPH, LogLogisticAFT, LogNormalAFT, WeibullAFT по метрикам CI, IBS_{RM}, AUPRC на всех наборах данных и по метрике IAUC на наборах данных ROTT2, WUHAN, SUPPORT2. Таким образом, предложенное дерево выживания с регуляризованными критериями превосходит существующие методы по совокупности метрик.

В табл. 6 представлены сводные результаты о качестве моделей по всем наборам данных. Каждая ячейка в таблице отражает средний ранг модели среди всех наборов данных, причем более низкий ранг

2024, **25** (3), 357–377. doi 10.26089/NumMet.v25r328

Таблица 4. Сравнение качества прогнозирования на наборе SMARTO

Table 4. Comparison of forecasting quality on the SMARTO dataset

Название модели Model	CI	IBS_{RM}	IAUC	AUPRC
CoxPH	0.4726	0.1692	**0.6463**	0.8494
KM	0.5000	0.1870	0.4999	0.8406
LogLogisticAFT	0.4700	0.1729	0.6453	0.8465
LogNormalAFT	0.4521	0.1804	0.6289	0.8400
WeibullAFT	0.5018	0.1692	**0.6605**	0.8483
ST	0.4711	0.1843	0.5717	0.8459
TREE($\lambda = 0$)	0.5329	0.1543	0.6189	0.8900
TREE($\lambda = 0.01$)	0.5348	0.1561	0.6140	0.8897
TREE($\lambda = 0.1$)	0.5485	0.1534	0.6190	**0.8905**
TREE($\lambda = 0.5$)	**0.5579**	**0.1533**	0.6192	**0.8905**
TREE($\lambda = 0.9$)	**0.5582**	**0.1525**	0.6184	0.8898

Таблица 5. Сравнение качества прогнозирования на наборе SUPPORT2

Table 5. Comparison of forecasting quality on the SUPPORT2 dataset

Название модели Model	CI	IBS_{RM}	IAUC	AUPRC
CoxPH	0.7860	0.1260	0.8623	0.4641
KM	0.5000	0.1777	0.4999	0.2780
LogLogisticAFT	0.7892	0.1088	0.8679	0.4692
LogNormalAFT	0.7891	0.1087	0.8703	0.4705
WeibullAFT	0.7888	0.1171	0.8708	0.4559
ST	0.7492	0.1240	0.8442	0.4550
TREE($\lambda = 0$)	**0.8000**	**0.1052**	**0.8893**	**0.5796**
TREE($\lambda = 0.01$)	**0.8023**	**0.1050**	**0.8911**	**0.5860**
TREE($\lambda = 0.1$)	0.7986	0.1055	0.8883	0.5771
TREE($\lambda = 0.5$)	0.7982	0.1065	0.8857	0.5750
TREE($\lambda = 0.9$)	0.7981	0.1060	0.8869	0.5757

Таблица 6. Средний ранг по качеству прогнозирования среди существующих и предлагаемых моделей по всем наборам данных

Table 6. Average rank of forecasting quality among existing and proposed models across all datasets

Название модели Model	CI	IBS_{RM}	IAUC	AUPRC
CoxPH	4.83	5.5	4.33	5.17
KM	6.67	6.83	7.33	6.83
LogLogisticAFT	4.83	5.17	4.5	5.5
LogNormalAFT	5.67	5.67	4.5	5.5
WeibullAFT	5.17	4.83	3.83	5.17
ST	6.33	6.0	6.17	5.83
TREE($\lambda = 0$)	3.0	2.5	3.0	2.0
TREE($\lambda = 0.01$)	2.33	2.17	2.83	**1.83**
TREE($\lambda = 0.1$)	**1.83**	2.17	2.83	**1.33**
TREE($\lambda = 0.5$)	2.0	**1.83**	**2.17**	2.0
TREE($\lambda = 0.9$)	**1.33**	**1.33**	**2.5**	2.83

2024, **25** (3), 357–377. doi 10.26089/NumMet.v25r328

указывает на лучшее относительное качество модели. Лучшие два значения по каждой метрике выделены жирным. Наименьший ранг показывает предложенная модель TREE, а лучшие значения достигаются при коэффициенте регуляризации 0.9. Исходя из результатов экспериментального исследования, можно сделать вывод, что использование регуляризации повысило гибкость и качество прогнозирования предложенного дерева выживания.

6. Заключение. Модели анализа выживаемости позволяют прогнозировать индивидуальное изменение вероятности наступления события во времени. Существующие методы имеют высокое качество прогнозирования, однако основаны на строгих предположениях. В частности, статистические модели предполагают неинформативность цензурирования — отсутствие связей между причиной потери наблюдения и проведением исследования. В реальных данных часто наблюдается информативность цензурирования, приводящая к появлению бимодального распределения времени событий. На данный момент в литературе не проводились исследования применимости моделей выживаемости к мультимодальным данным.

Критерий log-rank имеет низкую чувствительность к выборкам с мультимодальным распределением времени, если одна из выборок исчерпана или размеры выборок малы. Для преодоления данного недостатка в работе предложен метод регуляризации статистики log-rank, который учитывает информацию об априорном распределении времени событий при поиске разбиения и определяет ненулевой вклад всех интервалов времени. Сочетание взвешенных схем критерия log-rank и регуляризации позволяет учитывать различия на всей временно́й шкале и определять значимость определенных событий.

Результаты экспериментального исследования показывают, что предложенный подход обеспечивает прирост качества по метрикам CI, $\mathrm{IBS}_{\mathrm{RM}}$, IAUC, AUPRC на четырех наборах данных. Дерево выживания с регуляризованными критериями разбиения TREE превзошло по качеству прогнозирования существующие методы Каплана–Мейера, пропорциональных рисков Кокса, ускоренного времени отказа и дерева выживания с классическим критерием log-rank.

Список литературы

1. *Gilboa S., Pras Y., Mataraso A., et al.* Informative censoring of surrogate end-point data in phase 3 oncology trials // European Journal of Cancer. 2021. **153**. 190–202. doi 10.1016/j.ejca.2021.04.044.
2. *Turkson A.J., Ayiah-Mensah F., Nimoh V.* Handling censoring and censored data in survival analysis: a standalone systematic literature review // International Journal of Mathematics and Mathematical Sciences. 2021. **2021**, N 1. Article Number 9307475. doi 10.1155/2021/9307475.
3. *Templeton A.J., Amir E., Tannock I.F.* Informative censoring — a neglected cause of bias in oncology trials // Nature Reviews Clinical Oncology. 2020. **17**, N 6. 327–328. doi 10.1038/s41571-020-0368-0.
4. *Knaus W.A., Harrell F.E., Lynn J., et al.* The SUPPORT prognostic model: objective estimates of survival for seriously ill hospitalized adults // Annals of Internal Medicine. 1995. **122**, N 3. 191–203. doi 10.7326/0003-4819-122-3-199502010-00007.
5. *Yan L., Zhang H.-T., Goncalves J., et al.* An interpretable mortality prediction model for COVID-19 patients // Nature Machine Intelligence. 2020. **2**, N 5. 283–288. doi 10.1038/s42256-020-0180-7.
6. *Royston P., Lambert P.C.* Flexible parametric survival analysis using Stata: beyond the Cox model. College Station: Stata Press, 2011.
7. *Castelijns M.C., Helmink M.A.G., Hageman S.H.J., et al.* Cohort profile: the Utrecht cardiovascular cohort–second manifestations of arterial disease (UCC-SMART) study–an ongoing prospective cohort study of patients at high cardiovascular risk in the Netherlands // BMJ Open. 2023. **13**, N 2. Article Number e066952. doi 10.1136/bmjopen-2022-066952.
8. *Hawkins D.M.* Quantile-quantile methodology - detailed results // https://arxiv.org/abs/2303.03215. Cited September 13, 2024.
9. *Nguyen H.D.* A two-sample Kolmogorov–Smirnov-like test for big data // Proc. Data Mining: 15th Australasian Conference (AusDM 2017), August 19–20, 2017, Melbourne, Australia. doi 10.1007/978-981-13-0292-3_6. https://espace.library.uq.edu.au/view/UQ:f921d22. Cited September 13, 2024.
10. *Kaplan E.L., Meier P.* Nonparametric estimation from incomplete observations // Journal of the American Statistical Association. 1958. **53**, N 282. 457–481. doi 10.2307/2281868.

11. *Aalen O.O., Borgan O., Gjessing H.K.* Survival and event history analysis: a process point of view. New York: Springer, 2008. doi 10.1007/978-0-387-68560-1.
12. *Cox D.R.* Regression models and life-tables // Journal of the Royal Statistical Society: Series B Methodological. 1972. **34**, N 2. 187–202. doi 10.1111/j.2517-6161.1972.tb00899.x.
13. *Wei L.J.* The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis // Statistics in Medicine. 1992. **11**, N 14–15. 1871–1879. doi 10.1002/sim.4780111409.
14. *Shimokawa A., Kawasaki Y., Miyaoka E.* Comparison of splitting methods on survival tree // The International Journal of Biostatistics. 2015. **11**, N 1. 175–188. doi 10.1515/ijb-2014-0029.
15. *Gordon L., Olshen R.A.* Tree-structured survival analysis // Cancer Treatment Reports. 1985. **69**, N 10. 1065–1069.
16. *Lee S.-H.* Weighted log-rank statistics for accelerated failure time model // Stats. 2021. **4**, N 2. 348–358. doi 10.3390/stats4020023.
17. *Buyske S., Fagerstrom R., Ying Z.* A class of weighted log-rank tests for survival data when the event is rare // Journal of the American Statistical Association. 2000. **95**, N 449. 249–258. doi 10.1080/01621459.2000.10473918.
18. *Kotsiantis S.B.* Decision trees: a recent overview // Artificial Intelligence Review. 2013. **39**, N 4. 261–283. doi 10.1007/s10462-011-9272-4.
19. *Leblanc M., Crowley J.* Survival trees by goodness of split // Journal of the American Statistical Association. 1993. **88**, N 422. 457–467. doi 10.1080/01621459.1993.10476296.
20. *Costa V.G., Pedreira C.E.* Recent advances in decision trees: an updated survey // Artificial Intelligence Review. 2023. **56**, N 5. 4765–4800. doi 10.1007/s10462-022-10275-5.
21. *Harrell F.E., Lee K.L., Mark D.B.* Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors // Statistics in Medicine. 1996. **15**, N 4. 361–387. doi 10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4.
22. *Heagerty P.J., Zheng Y.* Survival model predictive accuracy and ROC curves // Biometrics. 2005. **61**, N 1. 92–105. doi 10.1111/j.0006-341X.2005.030814.x.
23. *Hung H., Chiang C.-T.* Estimation methods for time-dependent AUC models with survival data // Canadian Journal of Statistics. 2010. **38**, N 1. 8–26. doi 10.1002/cjs.10046.
24. *Lambert J., Chevret S.* Summary measure of discrimination in survival models based on cumulative/dynamic time-dependent ROC curves // Statistical Methods in Medical Research. 2016. **25**, N 5. 2088–2102. doi 10.1177/0962280213515571.
25. *Vasilev I., Petrovskiy M., Mashechkin I.* Sensitivity of survival analysis metrics // Mathematics. 2023. **11**, N 20. Article Number 4246. doi 10.3390/math11204246.
26. *Murphy A.H.* A new vector partition of the probability score // Journal of Applied Meteorology and Climatology. 1973. **12**, N 4. 595–600. doi 10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2.
27. *Haider H., Hoehn B., Davis S., Greiner R.* Effective ways to build and evaluate individual survival distributions // J. Mach. Learn. Res. 2020. **21**, N 1 Article Number 85, 3289–3351.
28. *Avati A., Duan T., Zhou S., et al.* Countdown regression: sharp and calibrated survival predictions // Proc. 35th Uncertainty in Artificial Intelligence Conf. PMLR, 2020. https://proceedings.mlr.press/v115/avati20a.html. Cited September 13, 2024.
29. *Fleming T.R., Harrington D.P., O'sullivan M.* Supremum versions of the log-rank and generalized Wilcoxon statistics // Journal of the American Statistical Association. 1987. **82**, N 397. 312–320. doi 10.1080/01621459.1987.10478435.
30. *Lee S.-H.* On the versatility of the combination of the weighted log-rank statistics // Computational Statistics & Data Analysis. 2007. **51**, N 12. 6557–6564. doi 10.1016/j.csda.2007.03.006.
31. *Vasilev I., Petrovskiy M., Mashechkin I.* Survival analysis algorithms based on decision trees with weighted log-rank criteria // Proc. 11th Int. Conf. on Pattern Recognition Applications and Methods (ICPRAM), Online, February 3–5, 2022. doi 10.5220/0000155500003122.
32. *Good I.J.* Weight of evidence: a brief survey // Bayesian statistics. 1985. **2**. 249–270.
33. *Napierala M.A.* What is the Bonferroni correction? // https://docs.ufpr.br/~giolo/LivroADC/Material/S3_Bonferroni%20Correction.pdf. Cited September 15, 2024.
34. *Refaeilzadeh P., Tang L., Liu H.* Cross-validation // Encyclopedia of database systems. Boston: Springer, 2009. 532–538. doi 10.1007/978-0-387-39940-9_565.
35. *Pölsterl S.* scikit-survival: a library for time-to-event analysis built on top of scikit-learn // J. Mach. Learn. Res. 2020. **21**, N 1. Article Number 212, 8747–8752.

2024, **25** (3), 357–377. doi 10.26089/NumMet.v25r328

36. *Davidson-Pilon C.* Lifelines: survival analysis in Python // Journal of Open Source Software. 2019. **4**, N 40. Article Number 1317. doi 10.21105/joss.01317.

37. *Васильев Ю.А.* Разработка библиотеки древовидных моделей анализа выживаемости // Вестник Московского университета. Серия 15: Вычислительная математика и кибернетика. 2024. № 3. 60–72.

Поступила в редакцию
4 сентября 2024 г.

Принята к публикации
5 сентября 2024 г.

Информация об авторах

Юлий Алексеевич Васильев — математик первой категории; Московский государственный университет имени М. В. Ломоносова, факультет вычислительной математики и кибернетики, кафедра интеллектуальных информационных технологий, Ленинские горы, 1, стр. 52, 119234, Москва, Российская Федерация.

Михаил Игоревич Петровский — к.ф.-м.н., доцент; Московский государственный университет имени М. В. Ломоносова, факультет вычислительной математики и кибернетики, кафедра интеллектуальных информационных технологий, Ленинские горы, 1, стр. 52, 119234, Москва, Российская Федерация.

Игорь Валерьевич Машечкин — д.ф.-м.н., профессор; Московский государственный университет имени М. В. Ломоносова, факультет вычислительной математики и кибернетики, кафедра интеллектуальных информационных технологий, Ленинские горы, 1, стр. 52, 119234, Москва, Российская Федерация.

References

1. S. Gilboa, Y. Pras, A. Mataraso, et al., "Informative Censoring of Surrogate End-Point Data in Phase 3 Oncology Trials," Eur. J. Cancer **153**, 190–202 (2021). doi 10.1016/j.ejca.2021.04.044.
2. A. J. Turkson, F. Ayiah-Mensah, and V. Nimoh, "Handling Censoring and Censored Data in Survival Analysis: A Standalone Systematic Literature Review," Int. J. Math. Math. Sci. **2021** (1), Article Number 9307475 (2021). doi 10.1155/2021/9307475.
3. A. J. Templeton, E. Amir, and I. F. Tannock, "Informative Censoring — a Neglected Cause of Bias in Oncology Trials," Nat. Rev. Clin. Oncol. **17** (6), 327–328 (2020). doi 10.1038/s41571-020-0368-0.
4. W. A. Knaus, F. E. Harrell, J. Lynn, et al., "The SUPPORT Prognostic Model: Objective Estimates of Survival for Seriously Ill Hospitalized Adults," Ann. Intern. Med. **122** (3), 191–203 (1995). doi 10.7326/0003-4819-122-3-199502010-00007.
5. L. Yan, H.-T. Zhang, J. Goncalves, et al., "An Interpretable Mortality Prediction Model for COVID-19 Patients," Nat. Mach. Intell. **2** (5), 283–288 (2020). doi 10.1038/s42256-020-0180-7.
6. P. Royston and P. C. Lambert, *Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model* (Stata Press, College Station, 2011).
7. M. C. Castelijns, M. A. G. Helmink, S. H. J. Hageman, et al., "Cohort Profile: the Utrecht Cardiovascular Cohort–Second Manifestations of Arterial Disease (UCC-SMART) Study–an Ongoing Prospective Cohort Study of Patients at High Cardiovascular Risk in the Netherlands," BMJ Open **13** (2), Article Number e066952 (2023). doi 10.1136/bmjopen-2022-066952.
8. D. M. Hawkins, "Quantile-Quantile Methodology – Detailed Results," https://arxiv.org/abs/2303.03215. Cited September 13, 2024.
9. H. D. Nguyen, "A Two-Sample Kolmogorov–Smirnov-like Test for Big Data," in Proc. Data Mining: 15th Australasian Conf. (AusDM 2017), Melbourne, Australia, August 19–20, 2017. doi 10.1007/978-981-13-0292-3_6. https://espace.library.uq.edu.au/view/UQ:f921d22. Cited September 13, 2024.
10. E. L. Kaplan and P. Meier, "Nonparametric Estimation from Incomplete Observations," J. Am. Stat. Assoc. **53** (282), 457–481 (1958). doi 10.2307/2281868.

2024, **25** (3), 357–377. doi 10.26089/NumMet.v25r328

11. O. O. Aalen, O. Borgan, and H. K. Gjessing, *Survival and Event History Analysis: A Process Point of View* (Springer, New York, 2008). doi 10.1007/978-0-387-68560-1.
12. D. R. Cox, "Regression Models and Life-Tables," J. R. Stat. Soc. Ser. B Methodol. **34** (2), 187–202 (1972). doi 10.1111/j.2517-6161.1972.tb00899.x.
13. L. J. Wei, "The Accelerated Failure Time Model: A Useful Alternative to the Cox Regression Model in Survival Analysis," Stat. Med. **11** (14–15), 1871–1879 (1992). doi 10.1002/sim.4780111409.
14. A. Shimokawa, Y. Kawasaki, and E. Miyaoka, "Comparison of Splitting Methods on Survival Tree," Int. J. Biostat. **11** (1), 175–188 (2015). doi 10.1515/ijb-2014-0029.
15. L. Gordon and R. A. Olshen, "Tree-Structured Survival Analysis," Cancer Treat. Rep. **69** (10), 1065–1069 (1985).
16. S.-H. Lee, "Weighted Log-Rank Statistics for Accelerated Failure Time Model," Stats **4** (2), 348–358 (2021). doi 10.3390/stats4020023.
17. S. Buyske, R. Fagerstrom, and Z. Ying, "A Class of Weighted Log-Rank Tests for Survival Data when the Event is Rare," J. Am. Stat. Assoc. **95** (449), 249–258 (2000). doi 10.1080/01621459.2000.10473918.
18. S. B. Kotsiantis, "Decision Trees: A Recent Overview," Artif. Intell. Rev. **39** (4), 261–283 (2013). doi 10.1007/s10462-011-9272-4.
19. M. Leblanc and J. Crowley, "Survival Trees by Goodness of Split," J. Am. Stat. Assoc. **88** (422), 457–467 (1993). doi 10.1080/01621459.1993.10476296.
20. V. G. Costa and C. E. Pedreira, "Recent Advances in Decision Trees: An Updated Survey," Artif. Intell. Rev. **56** (5), 4765–4800 (2023). doi 10.1007/s10462-022-10275-5.
21. F. E. Harrell, K. L. Lee, and D. B. Mark, "Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors," Stat. Med. **15** (4), 361–387 (1996). doi 10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4.
22. P. J. Heagerty and Y. Zheng, "Survival Model Predictive Accuracy and ROC Curves," Biometrics **61** (1), 92–105 (2005). doi 10.1111/j.0006-341X.2005.030814.x.
23. H. Hung and C.-T. Chiang, "Estimation Methods for Time-Dependent AUC Models with Survival Data," Can. J. Stat. **38** (1), 8–26 (2010). doi 10.1002/cjs.10046.
24. J. Lambert and S. Chevret, "Summary Measure of Discrimination in Survival Models Based on Cumulative/-Dynamic Time-Dependent ROC Curves," Stat. Methods Med. Res. **25** (5), 2088–2102 (2016). doi 10.1177/0962280213515571.
25. I. Vasilev, M. Petrovskiy, and I. Mashechkin, "Sensitivity of Survival Analysis Metrics," Mathematics **11** (20), Article Number 4246 (2023). doi 10.3390/math11204246.
26. A. H. Murphy, "A New Vector Partition of the Probability Score," J. Appl. Meteorol. Climatol. **12** (4), 595–600 (1973). doi 10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2.
27. H. Haider, B. Hoehn, S. Davis, and R. Greiner, "Effective Ways to Build and Evaluate Individual Survival Distributions," J. Mach. Learn. Res. **21** (1), Article Number 85, 3289–3351 (2020).
28. A. Avati, T. Duan, S. Zhou, et al., "Countdown Regression: Sharp and Calibrated Survival Predictions," in *Proc. 35th Uncertainty in Artificial Intelligence Conf. PMLR, 2020*. https://proceedings.mlr.press/v115/avati20a.html. Cited September 13, 2024.
29. T. R. Fleming, D. P. Harrington, and M. O'sullivan, "Supremum Versions of the Log-Rank and Generalized Wilcoxon Statistics," J. Am. Stat. Assoc. **82** (397), 312–320 (1987). doi 10.1080/01621459.1987.10478435.
30. S.-H. Lee, "On the Versatility of the Combination of the Weighted Log-Rank Statistics," Comput. Stat. Data Anal. **51** (12), 6557–6564 (2007). doi 10.1016/j.csda.2007.03.006.
31. I. Vasilev, M. Petrovskiy, and I. Mashechkin, "Survival Analysis Algorithms Based on Decision Trees with Weighted Log-Rank Criteria," in *Proc. 11th Int. Conf. on Pattern Recognition Applications and Methods (ICPRAM), Online, February 3–5, 2022*. doi 10.5220/0000155500003122.
32. I. J. Good, "Weight of Evidence: A Brief Survey," Bayesian Stat. **2**, 249–270 (1985).
33. "What is the Bonferroni Correction?" https://docs.ufpr.br/~giolo/LivroADC/Material/S3_Bonferroni%20Correction.pdf. Cited September 15, 2024.
34. P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-Validation," Encyclopedia of Database Systems (Springer, Boston, 2009), pp. 532–538. doi 10.1007/978-0-387-39940-9_565.
35. S. Pölsterl, "scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn," J. Mach. Learn. Res. **21** (1), Article Number 212, 8747–8752 (2020).

2024, **25** (3), 357–377. doi 10.26089/NumMet.v25r328

36. C. Davidson-Pilon, "Lifelines: Survival Analysis in Python," J. Open Source Softw. **4** (40), Article Number 1317 (2019). doi 10.21105/joss.01317.

37. I. A. Vasilev, "Developing Library of Tree-Based Models for Survival Analysis," Vestn. Mosk. Univ., Ser. 15: Vychisl. Mat. Kibern., No. 3, 60–72 (2024).

Received
September 4, 2024

Accepted for publication
September 5, 2024

Information about the authors

Iulii A. Vasilev — first category mathematician; Lomonosov Moscow State University, Faculty of Computational Mathematics and Cybernetics, Department of Intelligent Information Technologies, Leninskie Gory, 1, building 52, 119234, Moscow, Russia.

Mikhail I. Petrovskiy — Ph. D., associate professor; Lomonosov Moscow State University, Faculty of Computational Mathematics and Cybernetics, Department of Intelligent Information Technologies, Leninskie Gory, 1, building 52, 119234, Moscow, Russia.

Igor V. Mashechkin — Dr. Sci., professor; Lomonosov Moscow State University, Faculty of Computational Mathematics and Cybernetics, Department of Intelligent Information Technologies, Leninskie Gory, 1, building 52, 119234, Moscow, Russia.