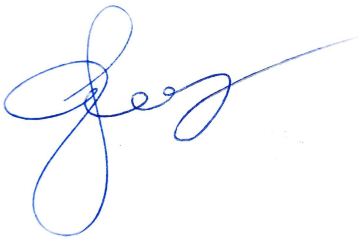
МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
имени М.В. ЛОМОНОСОВА

На правах рукописи



Васильев Юлий Алексеевич

ИССЛЕДОВАНИЕ И РАЗРАБОТКА МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ АНАЛИЗА ВЫЖИВАЕМОСТИ

Специальность 2.3.5 —
«Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей»

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата физико-математических наук

Москва – 2024

Работа выполнена на кафедре Интеллектуальных информационных технологий факультета Вычислительной математики и кибернетики Московского государственного университета им. М.В. Ломоносова.

Научный руководитель: **Петровский Михаил Игоревич**, кандидат физико-математических наук, доцент

Официальные оппоненты: **Шевцова Ирина Геннадьевна,** доктор физико-математических наук, кафедра Математической статистики факультета Вычислительной математики и кибернетики МГУ имени М.В. Ломоносова, профессор

Дьяконов Александр Геннадьевич, доктор физико-математических наук, доцент, Акционерное общество «ТБанк», академический руководитель направления наук о данных

Сотников Александр Николаевич, доктор физико-математических наук, профессор, Федеральное государственное бюджетное учреждение «Национальный исследовательский центр «Курчатовский институт», Отделение суперкомпьютерных систем и параллельных вычислений, главный научный сотрудник

Защита диссертации состоится «26» декабря 2024 г. в 14:00 часов на заседании диссертационного совета МГУ.012.2 Московского государственного университета имени М.В. Ломоносова по адресу: 119991, Москва, ГСП-1, Ленинские горы, МГУ, д. 1 строение 52, факультет Вычислительной математики и кибернетики, аудитория №238.

E-mail : ds012.2@cs.msu.ru

С диссертацией можно ознакомиться в отделе диссертаций научной библиотеки МГУ имени М.В. Ломоносова (Ломоносовский проспект, д.27) и на портале: `https://dissovet.msu.ru/dissertation/3259`

Автореферат разослан «____» ноября 2024 года.

Ученый секретарь
диссертационного совета МГУ.012.2,
кандидат физико-математических наук — Антонов А.С.

Общая характеристика работы

Актуальность темы исследования

Интеллектуальные системы анализа событий широко используются в медицине, биостатистике, социологии и анализе технологических процессов. Например, в медицине прогнозируется ожидаемое время и вероятность летального исхода, а в анализе надежности время технического сбоя или поломки оборудования. Интеллектуальные модели позволяют описывать контекст события, интерпретировать зависимости и прогнозировать время наступления события на основе характеристик объектов исследования (наблюдений).

Модели анализа выживаемости позволяют оценивать вероятность и время до наступления определенного события. Для сбора данных определяется целевое событие и фиксируется интервал исследования, в рамках которого могут появляться новые наблюдения. Каждому наблюдению сопоставляется вектор признаков X, полученный на момент начала исследования, а также время наступления события T.

Наблюдения, для которых наступает целевое событие, называются терминальными. Однако, полные данные могут быть недоступны и истинное время наступления события неизвестно в случае потери наблюдения или раннего прекращения исследования. Наблюдения с неизвестным временем события называются цензурированными. Например, в исследованиях летального исхода причиной цензурирования может быть перевод пациентов в другое учреждение, выписка или отказ пациента от исследования. Важно отметить, что наиболее распространено правое цензурирование, при котором известен момент выхода из исследования до наступления целевого события. Таким образом, уникальность анализа выживаемости заключается в использовании двух целевых переменных: времени события T и флага цензурирования δ.

Особенностью моделей анализа выживаемости является возможность прогнозирования функций вероятности наступления события для каждого момента времени. Функция выживания (survival function) определяет вероятность ненаступления события по истечении определенного времени $S(t) = P(T \geq t)$, где t – время наблюдения, T – случайная величина времени события. Функция плотности (density function) определяет риск наступления события $f(t) = (1 - S(t))^{\emptyset}$ в момент времени t. Функция риска (hazard function) определяет относительный риск события $h(t) = f(t)/S(t)$ в момент времени t при условии, что событие не наступило ранее. Системы интеллектуального

анализа событий должны обеспечивать прогноз данных функций в зависимости от характеристик наблюдения для каждого момента времени.

Построение прикладных интеллектуальных систем анализа событий напрямую связано со следующими особенностями реальных данных:

- Гетерогенность признакового пространства. Для описания состояния наблюдения используются непрерывные и категориальные показатели, которые могут содержать пропущенные значения из-за ограниченности информации или наличия ошибок;
- Распределение вероятностей времени событий. Постановка задачи и формат исследования влияют на распределение вероятностей времени и соотношение терминальных и цензурированных событий;
- Информативность цензурирования. Если причина цензурирования не связана с условиями проведения исследования, то говорят о неинформативном цензурировании, в противном случае существуют неучтенные факторы и цензурирование считается информативным.

Классические модели анализа выживаемости не позволяют работать с представленными особенностями данных и используют строгие предположения. Таким образом, актуальным является разработка интеллектуальных систем анализа выживаемости, не использующих строгие статистические предположения и применимых к особенностям реальных данных.

Степень разработанности темы

Построение интеллектуальных систем анализа выживаемости является перспективным направлением исследований и применяется в здравоохранении, анализе надежности и биостатистике. Большинство исследований посвящены применению классических статистических подходов и методов машинного обучения для анализа событий. Существующие решения основаны на следующих концепциях:

- Непараметрические методы не учитывают связь между признаками наблюдения и целевыми переменными и предполагают неинформативность цензурирования. Полупараметрические методы основаны на идее масштабирования непараметрической функции риска по индивидуальному для каждого наблюдения коэффициенту масштабирования. Параметрические методы предполагают теоретическое распределение времени, описывая индивидуальный прогноз как сдвиг функции во времени.
- Дискретные модели машинного обучения прогнозируют вектор вероятностей наступления события в фиксированные моменты времени. Регрессионные методы прогнозируют одну целевую переменную, но учи-

тывают полную информацию при расчете функции потерь. Ансамбли регрессионных моделей строят отдельную модель для каждого момента времени. Нейросетевые модели устанавливают размер выходного слоя количеству точек фиксированной временной шкалы, минимизируя отклонения между прогнозом и теоретической дискретной функцией.

- Непрерывные модели машинного обучения основаны на расширении статистических моделей. Регрессионные модели масштабируют базовую функцию на основе точечного прогноза относительного риска события. Модели деревьев выживаемости рекурсивно разбивают признаковое пространство по статистическому критерию на группы с максимально различной выживаемостью. Модели ансамблирования деревьев агрегируют прогнозы множества моделей, повышая качество прогнозирования, но теряя интерпретируемость. В таком случае строгость предположений зависит от критерия разбиения и непараметрических оценок в листах дерева.
- Для оценки качества прогнозирования используются точечные и интегральные метрики. Точечные метрики основаны на сравнении ожидаемой вероятности и времени события, а также единичных значений функций. Интегральные метрики оценивают значения функций для всех моментов времени путем сравнения с эталонной функцией или ранжирования наблюдений по риску наступления события. Наибольшую популярность получили метрики: правдоподобие, индекс согласованности и интегральный показатель Браера.

Существующие модели анализа выживаемости обладают несколькими недостатками. Статистические модели основаны на строгих предположениях, которые могут не выполняться на реальных данных. Дискретные модели прогнозируют значения функций в рамках ограниченной временной шкалы. Непрерывные модели используют статистические предположения для прогнозирования функций. Важно отметить, что существующие модели анализа выживаемости не позволяют непосредственно обрабатывать категориальные и пропущенные значения и требуют предварительной обработки данных.

Возможным путем преодоления существующих недостатков является разработка подхода построения моделей анализа выживаемости на основе деревьев решений, поскольку в задачах машинного обучения они позволяют определять зависимости без необходимости предварительного определения предположений модели и обработки данных. Модификация этапа поиска разбиения и построения листовых оценок позволит обрабатывать катего-

риальные и пропущенные значения для обучения и применения модели, а также преодолеть строгие предположения критериев разбиения и непараметрических оценок. Также, необходимо разработать программную библиотеку анализа выживаемости.

Целью диссертационной работы является разработка математического и программного обеспечения интеллектуальной системы для решения задач анализа выживаемости с использованием методов машинного обучения на основе деревьев решений.

Объектом исследования диссертационной работы являются модели анализа выживаемости, позволяющие прогнозировать время и вероятность наступления события, а также функции выживания и риска. Предметом исследования диссертационной работы является разработка алгоритмов построения моделей анализа выживаемости, применимых к неполным непрерывным и категориальным данным, а также к случаям информативного цензурирования без использования строгих статистических предположений.

Для достижения цели необходимо решение следующих задач:

1. Разработать методы построения интерпретируемых моделей анализа выживаемости на основе деревьев решений, учитывающих особенности реальных данных.
2. Исследовать и разработать методы оценки качества прогнозирования моделей анализа выживаемости;
3. Разработать алгоритмы ансамблирования предложенных деревьев выживаемости, позволяющих повысить качество прогнозирования;
4. Реализовать интеллектуальную программную систему на основе разработанного комплекса алгоритмов анализа выживаемости и провести её апробацию на прикладной задаче анализа медицинских данных.

Диссертация соответствует специальности 2.3.5 «Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей» в части направления разработки интеллектуальных систем машинного обучения и инструментальных средств разработки цифровых продуктов, поскольку целью работы является исследование, разработка и программная реализация комплекса алгоритмов для построения интеллектуальной системы анализа событий, применимой для решения задач анализа выживаемости.

Необходимость работы с реальными данными низкого качества существенно ограничивает возможность применения в таких системах классических моделей, использующих статистические подходы. Для решения этой проблемы необходимо разработать новые оригинальные методы или пред-

ложить модификации существующих методов машинного обучения, а также реализовать их в виде программной библиотеки с открытым кодом, которая может быть использована для построение интеллектуальных систем анализа событий для широкого спектра прикладных областей. В рамках настоящей работы библиотека будет использоваться для решения ряда прикладных задач анализа выживаемости из области медицины.

Научная новизна

Разработан алгоритм поиска разбиений в данных с цензурированием, основанный на гистограммном вычислении взвешенных критериев log-rank и учитывающий категориальные и пропущенные значения. Предложенный подход регуляризации критерия позволяет обрабатывать случаи информативного цензурирования, учитывая информацию об априорном распределении событий, в том числе в случае малых выборок, когда в процессе построения дерева возникает разреженная область в пространстве признаков. На основе алгоритма поиска разбиений предложен метод построения интерпретируемых деревьев выживаемости с модифицированными непараметрическими оценками функций выживания и риска. Для оценки качества прогнозирования исследованы существующие и предложены модифицированные метрики с равным вкладом событий и временных интервалов. Также предложены методы ансамблирования деревьев выживаемости, основанные на построении независимых базовых бутстеп моделей, а также на подходе усиления слабых моделей с использованием адаптивного бустинга с перевыборкой.

Теоретическая и практическая значимость

Разработанная программная библиотека анализа выживаемости призвана упростить процесс построения и применения моделей анализа выживаемости, оценки качества прогнозирования и проведения экспериментального исследования. Разработанные методы построения моделей могут использоваться для решения различных прикладных задач, основанных на анализе выживаемости. Апробация библиотеки проводилась на прикладных задачах анализа медицинских данных.

Комплекс предложенных алгоритмов позволяет строить модели анализа выживаемости, применимые к реальным данным. Метод построения деревьев выживаемости позволяет строить интерпретируемые прогнозы, а ансамбли деревьев имеют высокое качество прогнозирования. По результатам экспериментального исследования, предложенные методы превзошли по качеству существующие методы анализа выживаемости. Полученные результаты диссертационной работы могут послужить основой для построения перспектив-

ных современных систем анализа событий, которые будут включать в себя средства анализа выживаемости наблюдений. При этом, могут использоваться как все разработанные модули, так и отдельные из них.

Методология и методы исследования

При получении основных результатов диссертационной работы использовались методы машинного обучения и математической статистики. При разработке модулей программной библиотеки анализа выживаемости использовались методы объектно-ориентированного проектирования, а также методы векторизации и параллелизации вычислений.

Основные положения, выносимые на защиту:

1. Предложенный метод построения деревьев выживаемости, учитывающий особенности реальных данных: наличие категориальных признаков и пропущенных значений, распределения вероятностей времени наступления событий и информативность цензурирования. Алгоритм поиска разбиений в данных с цензурированием основан на взвешенных регуляризованных критериях log-rank и реализован в виде гистограммного метода. Метод позволяет строить интерпретируемые прогнозы времени и вероятности события, функций выживания и риска;
2. Предложенные методы построения бутстреп и бустинг ансамблей деревьев выживания позволяют достичь высокого качества прогнозирования за счет использования независимой и адаптивной схем агрегации прогнозов базовых моделей, формирующих ансамбль. В качестве функций потерь используются модифицированные метрики, которые обеспечивают равенство вкладов относительно целевого времени события, флага цензурирования и временной шкалы;
3. Разработанная программная библиотека `survivors` включает предложенный комплекс алгоритмов для построения интеллектуальных систем анализа выживаемости. По результатам экспериментального применения библиотеки на медицинских данных, предложенные методы превзошли по качеству прогнозирования существующие методы.

Достоверность полученных результатов обеспечивается проведенными экспериментами, открытым кодом реализованных методов и подходов, обоснованием принимаемых решений, публикациями в рецензируемых журналах и апробацией на российских и международных конференциях.

Апробация работы

Основные результаты работы докладывались на:

– Научная конференция «Тихоновские чтения» (Россия, Москва, 2021).

– 11th International Conference on Pattern Recognition Applications and Methods (Австрия, 2022).
– Научная конференция «Ломоносовские чтения» (Россия, Москва, 2022).
– XXIX Международная научная конференция студентов, аспирантов и молодых ученых «Ломоносов 2022» (Россия, Москва, 2022).
– Научная конференция «Ломоносовские чтения» (Россия, Москва, 2023).
– XXX Международная научная конференция студентов, аспирантов и молодых ученых «Ломоносов 2023» (Россия, Москва, 2023).
– IV кафедральная студенческая конференция «Artificial Intelligence and Creativity» (Россия, Москва, 2023).
– Научная конференция «Ломоносовские чтения» (Россия, Москва, 2024).
– Научный семинар МС ВМК МГУ под руководством В.Ю. Королева.

Результаты выступлений были изданы в 7 работах в сборниках тезисов и трудов конференций. Получено свидетельство о государственной регистрации программы для ЭВМ.

Результаты диссертационной работы использовались в следующих НИР:
– «Исследование, разработка и применение инновационных технологий построения интеллектуальных программных систем» (Номер договора: 6.2.18), 2018–2027 гг.
– «Выполнение работ в области разработки и внедрения методов искусственного интеллекта и анализа больших данных в сфере здравоохранения» (Номер договора: ЦАРСС-12/20-03/У), 2020–2021 гг.
– «Выполнение части работ по развитию прикладного программного обеспечения государственной информационной системы обязательного медицинского страхования» (Номер договора: № С/01-ПО7/02731000011210000030001), 2021–2022 гг.

Личный вклад автора заключается в выполнении основного объема теоретических и экспериментальных исследований, а также в разработке архитектуры и реализации открытой библиотеки анализа выживаемости. Подготовка части материалов к публикации проводилась совместно с соавторами, причем вклад диссертанта был определяющим. В работах [1, 2] М.И. Петровскому принадлежит постановка задачи применения моделей к категориальным и пропущенным значениям, а И.В. Машечкину принадлежат рекомендации к методологии исследований. В работе [4] М.И. Петровский и И.В. Машечкин участвовали в постановке задачи и анализе результатов. Создание программных реализаций алгоритмов и проведение всех численных экспериментов было выполнено автором полностью самостоятельно.

Диссертационное исследование является самостоятельным и законченным трудом автора.

Публикации

Основные результаты по теме диссертации изложены в 4 публикациях, изданных в рецензируемых научных изданиях, определенных в п. 2.3 Положения о присуждении ученых степеней в Московском государственном университете имени М.В. Ломоносова.

Объем и структура работы

Диссертационная работа состоит из введения, пяти глав, заключения и списка литературы. Полный объём диссертации составляет 142 страницы, включая 57 рисунков и 29 таблиц. Список литературы содержит 123 наименования.

Краткое содержание диссертации

Во **введении** описывается область исследования, обосновывается актуальность работы, раскрываются ее цели, обосновываются научная новизна, практическая и теоретическая значимость работы.

Первая глава посвящена обзору существующих методов анализа выживаемости, определяющих способ сбора данных, построение прогнозных моделей и оценку качества прогнозирования. Для каждого подхода описываются основные принципы и особенности, приводятся наболее известные алгоритмы, выделяются преимущества и недостатки.

Подраздел 1.1 посвящен обзору методов сбора данных с левым, правым и интервальным цензурированием, а также способам формализации задачи в зависимости от целей исследования. В данной работе рассматривается задача прогнозирования непрерывных функций анализа выживаемости при наличии правого цензурирования. Формально, анализируемые данные анализа выживаемости представляются в виде тройки значений (X_i, T_i, δ_i) для каждого наблюдения i. Вектор признаков X_i вычисляется при входе наблюдения в исследование. Целевыми переменными являются флаг цензурирования δ_i и время события T_i. К особенностям реальных данных относится: информативность цензурирования — наличие связи между потерей наблюдения и условиями исследования; гетерогенность признакового пространства; различное распределение вероятностей времени наступления событий.

В **подразделе 1.2** рассматриваются метрики оценки качества анализа выживаемости. Точечные метрики основаны на однократной оценке значений функций выживания и риска в момент наступления события. Интегральные

метрики основаны на сравнении прогнозируемой функции выживания и риска с целевой функцией. Метрика Concordance Index (CI) применяется для оценки времени наступления события T, метрики Integrated Brier Score (IBS) и Survival area under precision-recall curve (AUPRC) применяются для оценки функции выживания, а метрика Integrated area under ROC-curve (IAUC) применяется для оценки функции риска.

В подразделе 1.3 приводится обзор статистических методов анализа выживаемости, основанных на строгих предположениях. Непараметрические методы Каплана–Мейера (KM) и Нельсона–Аалена (NA) не описывают зависимость между признаками наблюдения и целевыми переменными, а также предполагают неинформативность флага цензурирования. Метод пропорциональных рисков Кокса (Cox PH) предполагает единую форму функции выживания и линейную зависимость между коэффициентом масштабирования и признаками наблюдения. Метод ускоренного времени отказа (AFT) предполагает теоретическое распределения времени события (Вейбулла, логнормальное, лог-логистическое). Применение статистических моделей ограничено, поскольку предположения могут не выполняться на практике.

Подраздел 1.4 посвящен обзору существующих методов анализа выживаемости на основе деревьев решений. Методы построения деревьев выживаемости основаны на рекурсивном разбиении пространства признаков на области, содержащие наблюдения с максимально близкой выживаемостью внутри области и максимально различной между областями. В листах деревьев выживаемости строится непараметрическая оценка времени наступления и вероятности события, а также функций выживания и риска. Разбиение производится на основе критерия log-rank, нулевая гипотеза которого предполагает равенство функций риска дочерних выборок. Таким образом, выбор разбиения с минимальным уровнем значимости нулевой гипотезы приводит к разделению пространства на области с различными функциями риска. Однако критерий разбиения и непараметрические модели основаны на предположении неинформативности цензурирования.

В подразделе 1.5 рассматриваются регрессионные методы машинного обучения, основанные на нейронных сетях, байесовских моделях и методе опорных векторов. Точечные методы прогнозируют только значения времени и вероятности события. Дискретные модели прогнозируют функции выживания и риска в рамках дискретной шкалы. Непрерывные модели сводят задачи прогнозирования функций анализа выживаемости к классификации и регрессии с последующим построением функций на основе статистических

моделей. Наиболее перспективными являются модели $DeepSurv$ и $CoxTime$, которые прогнозируют величину относительного риска на основе полносвязной нейронной сети. Модель $CoxTime$ учитывает влияние времени на относительный риск, преодолевая предположение пропорциональности рисков. Для прогнозирования функций выживания и риска, модели заменяют линейную комбинацию весов и признаков в модели Кокса на отклик нейронной сети.

В подразделе 1.6 рассматриваются методы ансамблирования моделей анализа выживаемости. Метод случайного леса основан на построении множества независимых деревьев выживания, а бустинг ансамбли основаны на усилении слабых статистических моделей.

Таким образом, существующие модели анализа выживаемости работают с заполненными числовыми данными и используют строгие статистические предположения. Наиболее перспективны методы на основе деревьев решений, позволяющие строить интерпретируемые прогнозы и использующие статистические предположения только на этапах поиска разбиения выборки и построения непараметрических оценок. Существующие метрики качества позволяют оценивать качество прогнозирования точечных величин и функций, однако не исследованы на предмет чувствительности к выделенным особенностям данных. На основе проведенного аналитического обзора формулируются направления дальнейших исследований:

- Исследование и разработка методов построения интерпретируемых деревьев выживаемости, применимых к категориальным и неполным данным с цензурированием и позволяющих прогнозировать время, вероятность события и непрерывные функции выживания и риска без учета статистических предположений;
- Исследование влияния особенностей данных на метрики качества анализа выживаемости и разработка метода оценки качества прогнозирования, определяющего равный вклад событий и временных интервалов;
- Исследование и разработка методов ансамблирования деревьев выживаемости для повышения качества прогнозирования без использования строгих статистических предположений.

Вторая глава посвящена исследованию и разработке методов построения деревьев выживаемости, применимых к категориальным и непрерывным данным, пропущенным значениям и случаям информативного цензурирования. Классический метод построения дерева выживаемости основан на рекурсивном разбиении признакового пространства на области с различной выживаемостью. Целью этапа поиска лучшего разбиения выборки с цензури-

рованием является выбор правила разбиения, максимизирующего различия между выживаемостью дочерних выборок. Каждому листовому узлу дерева сопоставляется непараметрическая оценка функции выживаемости и риска.

В подразделе 2.1 приводится описание используемых для исследования шести наборов медицинских данных. Наборы SUPPORT2, PBC, SMARTO содержат данные о летальных исходах среди онкологически больных пациентов с признаками анамнеза, статуса заболевания и клинических показателей. Наборы GBSG и ROTT2 содержат данные о рецидиве онкологических заболеваний у пациентов, которые описываются анамнезом, характеристиками опухоли и стратегии лечения. Набор WUHAN содержит данные о выписке пациентов с диагностированным заболеванием COVID-19 с признаками анамнеза и клинических показателей. Признаковое простанство наборов содержит категориальные признаки и пропущенные значения, а целевые переменные имеют различное распределение вероятностей времени событий и дисбаланс цензурирования. По результатам исследования, на рассматриваемых данных не выполняются статистические предположения пропорциональности рисков и неинформативности цензурирования.

В подразделе 2.2 предложен гистограммный метод поиска бинарного разбиения выборки цензурированных данных по непрерывному признаку f. Пусть каждое наблюдение представлено в виде тройки значений (v_k, T_k, δ_k), где T_k — время события, δ_k — флаг цензурирования, v_k — значение признака f. Определим v_i как упорядоченное множество значений признака f.

Для каждой промежуточной точки $s_i = \frac{v_i + v_{i-1}}{2}$ определяются две ветви разбиения: левая при $v \leq s_i$, правая при $v > s_i$. Для контроля вычислительной сложности, при числе промежуточных точек $n > 100$ производится квантилизация значений признака f. Пусть τ_j — упорядоченный набор времени наступления событий $\tau_1 < \tau_2 < \ldots < \tau_K$ в выборке. Тогда для каждого момента времени τ_j, количество наблюдений $n_{jjv=s}$ и событий $O_{jjv=s}$ в момент τ_j при условии $v = s$ равны:

$$n_{jjv=s} = \sum_k I(T_k = \tau_j) \cdot I(v_k = s),$$

$$O_{jjv=s} = \sum_k I(T_k = \tau_j) \cdot I(v_k = s) \cdot I(\delta_k = 1).$$

Для каждой промежуточной точки s_i определяются гистограммы левой и правой ветви: n^l и n^r (гистограммы наблюдений), O^l и O^r (гистограммы

событий) по следующим формулам:

$$n^l_{s_i,j} = \sum_{k:s_k \le s_i} n_{s_k,j},\ n^r_{s_i,j} = \sum_{k:s_k > s_i} n_{s_k,j}, \tag{1}$$

$$O^l_{s_i,j} = \sum_{k:s_k \le s_i} n_{s_k,j},\ O^r_{s_i,j} = \sum_{k:s_k > s_i} n_{s_k,j}. \tag{2}$$

Гистограммный метод основан на идее итеративного обновления гистограмм левой и правой ветви и выполняется по следующим правилам:

$$n^l_{0,j} = 0,\ n^r_{0,j} = n_j,$$

$$n^l_{s_i,j} = n^l_{s_{i-1},j} + n_{jjv=s_i},\ n^r_{s_i,j} = n^r_{s_{i-1},j} - n_{jjv=s_i},$$

$$O^l_{0,j} = 0,\ O^r_{0,j} = O_j,$$

$$O^l_{s_i,j} = O^l_{s_{i-1},j} + O_{jjv=s_i},\ O^r_{s_i,j} = O^r_{s_{i-1},j} - O_{jjv=s_i}.$$

Левые гистограммы $(n^l_{0,j}, O^l_{0,j})$ инициализируются 0 для всех моментов времени, а правые $(n^r_{0,j}, O^r_{0,j})$ — полным распределением времени наблюдений и событий. Итеративно исчерпывая упорядоченное множество точек s_i, для каждой промежуточной точки s_i вычисляются две гистограммы $(n_{jjv=s_i}, O_{jjv=s_i})$, которые векторно прибавляются к гистограмме левой ветви ($n^l_{s_i,j}$ и $O^l_{s_i,j}$ соответственно) и вычитаются из правой ($n^r_{s_i,j}$ и $O^r_{s_i,j}$ соответственно). Таким образом, для каждой точки s_i остаются верными равенства 1 и 2. На основе $n^l_{s_i,j}$ и $n^r_{s_i,j}$ вычисляются гистограммы оставшихся наблюдений $N^l_{s_i,j}, N^r_{s_i,j}$ и гистограмма ожидаемого числа событий в левой ветви $E^l_{s_i,j}$:

$$N^l_{s_i,j} = \sum_{\tau_k \ge \tau_j} n^l_{s_i,k},\ N^r_{s_i,j} = \sum_{\tau_k \ge \tau_j} n^r_{s_i,k},\ E^l_{s_i,j} = \frac{N^l_{s_i,j} \cdot O_j}{N_j}.$$

Используя гистограммы оставшихся наблюдений и наступивших событий, для каждой промежуточной точки s_j вычисляется значение взвешенной статистики log-rank:

$$LR_{s_i} = \frac{\sum\limits_{j=1}^{K} w_j \left(O^l_{s_i,j} - E^l_{s_i,j}\right)}{\sqrt{\sum\limits_{j=1}^{K} w_j^2 \cdot E^l_{s_i,j} \left(\frac{N_j - O_j}{N_j}\right) \left(\frac{N_j - N^l_{s_i,j}}{N_j - 1}\right)}}. \tag{3}$$

Для преодоления предположения равенства вероятности наступления события для всех моментов времени, предлагается использовать весовые схемы критерия log-rank: wilcoxon ($w_j = N_j$), tarone-ware ($w_j = \sqrt{N_j}$), peto-peto ($w_j = \hat{S}(\tau_j)$), где $\hat{S}(t)$ — оценка функции выживания по методу Каплана–Мейера. Квадрат взвешенной статистики log-rank имеет распределение хи-квадрат.

Лучшее разбиение для признака f определяется по максимальному значению статистики log-rank (или минимальному p-value) и соответствует паре правил $v \leq s_{best,f}$ и $v > s_{best,f}$, где $s_{best,f} = \arg\max_i LR_{s_i}$. Обработка категориальных значений производится путем отображения категорий на числовую шкалу по методу Weight Of Evidence (WoE). Метод основан на построении бинарной модели, описывающую связь между категориями признака и вероятностью наступления целевого события. Каждой категории B признака f сопоставляется значение $\ln \frac{P(DjB)}{P(\overline{D}jB)}$, где $P(DjB)$ – вероятность наступлении события для категории B, а $P(\overline{D}jB)$ – вероятность цензурировании события для категории B. Предложенный подход учитывает информативность цензурирования путем расположения категорий на основе схожести подвыборок по цензурируемости. Для обработки пропущенных значений рассматривается два варианта размещения пропусков в каждой из ветвей и определяется лучшая ветвь, при которой достигается наибольшая значимость разбиения.

В подразделе 2.3 предложен метод построения дерева выживания. Корневой узел содержит все исходные данные, а при разбиении узла для каждого допустимого признака определяется лучшее правило разбиения. При выборе лучшего признака разбиения применяется поправка Бонферрони на множественную проверку гипотез. Поправка заключается в расчете скорректированного значения p-value по количеству значимых разбиений и уменьшает значимость признаков в большим количеством точек разбиения. Лучшее правило разбиения определяется по минимальному значению p-value.

Описанный алгоритм приводит к разбиению корневого узла на две дочерние выборки. Далее, алгоритм разбиения выборок рекурсивно применятся для каждого дочернего узла. Для борьбы с переобучением модели разработаны подходы обрезки дерева путем контроля роста дерева (ограничение структуры на основе гиперпараметров) и удаления избыточных разбиений (выбор поддерева с лучшим качеством прогнозирования на валидационной выборке). При применении модели к внешним данным, каждому наблюдению сопоставляется листовой узел дерева на основе системы правил разбиения. К

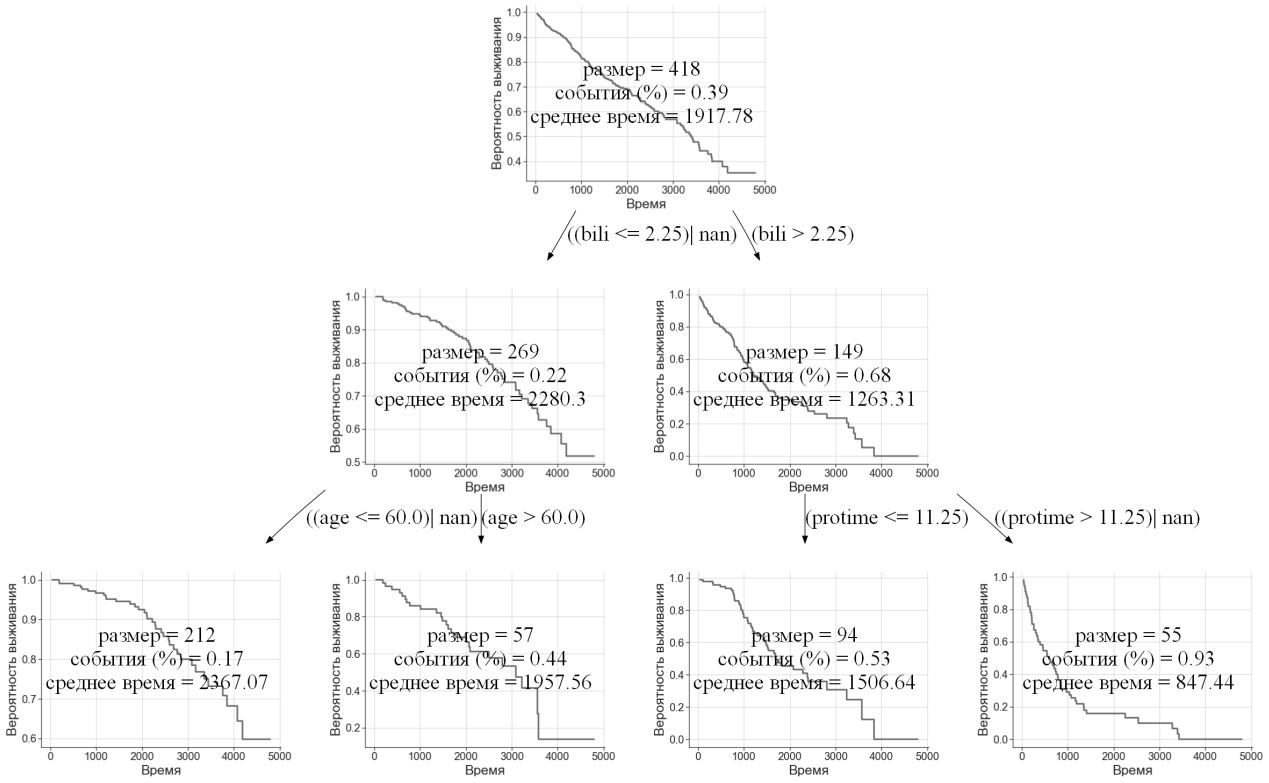


Рис. 1: Пример построенного дерева выживаемости глубины 2 на наборе PBC. Каждый узел дерева содержит визуализацию функции выживания — изменение вероятности выживания относительно времени, а также содержит информацию о размере листа (размер), доле терминальных событий (события (%)) и среднем времени события (среднее время).

непрерывным признакам применяются правила на основе интервалов, к категориальным признакам — правила на вхождение в множество допустимых категорий. Для пропущенных наблюдений выбирается ветвь, допускающая наличие пропусков.

В каждом листе дерева выживания строятся непараметрические оценки функции выживания и функции риска, а также вычисляются значения вероятности и ожидаемого времени наступления события. Таким образом, построенная модель позволяет прогнозировать вероятность и время наступления события, функции выживания и риска.

Пример построенного дерева глубины 2 на наборе PBC представлен на Рисунке 1. Дерево основано на признаках bili (показатель билирубина), protime (стандартизированное время свертывания крови), age (возраст). Значения «nan» указывают на ветвь определения пропущенных значений.

Подраздел 2.4 посвящен исследованию проблемы влияния информативности цензурирования на обучение дерева выживамости. Информатив-

ность цензурирования может приводить к появлению мультимодального распределения времени событий в узлах дерева выживания. Критерий log-rank не учитывает различия после исчерпания одной из выборок, а непараметрическая оценка Каплана–Мейера имеет низкое качество описания при работе с мультимодальными распределениями времени.

Для работы с данными с мультимодальным распределением времени предложен подход регуляризации критерия разбиения, который позволяет учитывать информацию об априорном распределении времени событий при поиске разбиений выборок. Подход основан на расчете в корневом узле дерева априорного количества оставшихся наблюдений N_j^A к моменту τ_j и априорного количества событий O_j^A в момент τ_j. При поиске лучшего разбиения узла, данная информация добавляется к текущему распределению времени событий и наблюдений с коэффициентом регуляризации λ. Формально, регуляризация критерия обеспечивается подстановкой в формулу (3) значений $\hat{N}_{1,j}, \hat{N}_{2,j}, \hat{O}_{1,j}, \hat{O}_{2,j}$, рассчитанных по формулам:

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

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

Для построения листовых оценок, устойчивых к случаям информативного цензурирования, предложена модификация оценки Каплана–Мейера. Непараметрическая модель $KMWV$ основана на генерации виртуальных событий по одномодальному распределению с последующим построением непараметрической оценки Каплана–Мейера. В частности, рассматривается семейство нормальных распределений $N(\overline{T}, S^2)$, и для каждой листовой выборки определяется теоретическое распределение на основе выборочного среднего $\overline{T} = \sum_{i=1}^{n} T_i/n$ и выборочной дисперсии $S^2 = \sum_{i=1}^{n}(T_i - \overline{T})^2/(n-1)$, где n — размер выборки, $\{T_i\}$ — множество значений времени в выборке.

Процесс порождения виртуальных событий основан на генерации времени $\tilde{T} \quad N(\overline{T}, S^2)$ с флагом цензурирования $\tilde{\delta} \quad Bernoulli((\sum_{i=1}^{n} \delta_i)/n)$, где $\{\delta_i\}$ — значения флага цензурирования в листовой выборке. Наконец, по множествам $\{\tilde{T}\}, \{\tilde{\delta}\}$ строится оценка Каплана–Мейера. Подход не требует выполнения предположения о неинформативности цензурирования и применим к данным с мультимодальным распределением времени событий.

Таким образом, предложенный алгоритм поиска разбиений выборки применим к гетерогенным данным (псевдокод алгоритма представлен на Рисунке 2). Подход отображения категориальных значений на числовую прямую по

Алгоритм 1 HistogramFindBestSplit – поиск лучшего разбиения выборки

Вход:

T ▷ Время наступления событий выборки
δ ▷ Флаг цензурирования наблюдений выборки
v ▷ Значение признака наблюдений выборки
T_A ▷ Время наступления событий корневой выборки
δ_A ▷ Флаг цензурирования наблюдений корневой выборки
λ ▷ Коэффициент регуляризации
$FeatureType$ ▷ Тип признака: непрерывный или категориальный
$Crit$ ▷ Весовая схема критерия log-rank

Выход: BestSplit ▷ Лучшее бинарное разбиение выборки по признаку

T_{nan}, δ_{nan} ← getNanValues(T, δ, v)
$T_{\overline{nan}}$, $\delta_{\overline{nan}}$, $v_{\overline{nan}}$ ← getNotNanValues(T, δ, v)
if FeatureType = Categorical **then**
    $v_{\overline{nan}}$, map_categ ← WeightOfEvidence($T_{\overline{nan}}$, $\delta_{\overline{nan}}$, $v_{\overline{nan}}$)
s_i ← GetQuantile($v_{\overline{nan}}$)

n_j^l, O_j^l ← 0
n_j^r, O_j^r ← CountHistogram($T_{\overline{nan}}$, $\delta_{\overline{nan}}$)
$n_{nan,j}$, $O_{nan,j}$ ← CountHistogram($T_{\overline{nan}}$, $\delta_{\overline{nan}}$)
n_j^A, O_j^A ← CountHistogram(T_A, δ_A)
(n_j^l, O_j^l) += $(\lambda * n_j^A, \lambda * O_j^A)$
(n_j^r, O_j^r) += $(\lambda * n_j^A, \lambda * O_j^A)$

Splits ← []
for s in s_i **do**
    Mask ← $v_{\overline{nan}} = s$
    $n_{j|v=s}$, $O_{j|v=s}$ ← CountHistogram($T_{\overline{nan}}$[Mask], $\delta_{\overline{nan}}$[Mask])
    (n_j^l, O_j^l) += $(n_{j|v=s}, O_{j|v=s})$
    (n_j^r, O_j^r) −= $(n_{j|v=s}, O_{j|v=s})$
    Splits ← Splits ∪ LogRankCriterion(n_j^l, O_j^l, $n_j^r + n_{nan,j}$, $O_j^r + O_{nan,j}$, $Crit$)
    Splits ← Splits ∪ LogRankCriterion($n_j^l + n_{nan,j}$, $O_j^l + O_{nan,j}$, n_j^r, O_j^r, $Crit$)
BestSplit ← SelectBest(Splits)
return BestSplit

Рис. 2: Псевдокод алгоритма поиска лучшего разбиения выборки.

методу Weight Of Evidence располагает категории на основе схожести подвыборок по цензурируемости. Применимость к случаям информативности цензурирования обеспечивается за счет подхода регуляризации, а высокая чувствительность к распределению вероятностей времени событий за счет взвешенных критериев log-rank. Предложенный метод построения деревьев выживаемости основан на алгоритме поиска разбиений и не использует статистических предположений. Для обработки информативности цензурирования

предложена модификация оценки Каплана–Мейера, используемая в листовых узлах дерева. Прогноз модели интерпретируем, поскольку для каждого листа соответствует набор правил, получаемый при проходе от корня к листу.

Третья глава посвящена исследованию влияния особенностей данных на метрики анализа выживаемости, разработке модификаций метрик и оценке качества предложенного метода построения деревьев выживания.

Подраздел 3.1 посвящен анализу чувствительности существующих метрик. Под избыточной чувствительностью метрики будем понимать наличие неравного вклада отдельных величин при равенстве условий оценки. Исследование чувствительности проводится в рамках случая неопределенности времени наступления события при $S(t) = 0.5$, поскольку такой прогноз не зависит от характеристик событий и интервалов временой шкалы. Рассматриваются 4 случая избыточной чувствительности метрик к вкладу отдельных событий, временных компонент, временных интервалов, дисбалансу цензурирования. Сформулированы и доказаны утверждения наличия избыточной чувствительности метрики IBS, экспериментально продемонстрированы случаи изыбыточной чувствительности метрик $IAUC, AUPRC$.

Утверждение 3. *Вклад отдельных наблюдений в метрике IBS монотонно неубывает относительно истинного времени события в случае неопределенности времени наступления события.*

Утверждение 4. *Вклад отдельных наблюдений в метрике $AUPRC$ не зависит от истинного времени события в случае неопределенности времени наступления события.*

Утверждение 5. *Метрика IBS определяет различные значения компонент $BS(t)$ в случае неопределенности времени наступления события.*

Утверждение 6. *Метрика IBS вогнута вниз относительно истинного времени наступления событий при равном прогнозе функции выживания.*

Для обеспечения равного вклада отдельных событий и временных компонент предлагается определять равный вес для наблюдений и использовать контролируемое усреднение вкладов во времени по оставшимся наблюдениям. Для учета всех временных интервалов при интегрировании метрики предлагается проводить интегрирование напрямую по времени, без использования весовых схем. Для определения равного вклада цензурированных и терминальных событий предлагается метрика со сбалансированным усреднением значений по каждому классу во времени.

Разработаны модификации для преодоления избыточной чувствительности метрик к выявленным случаям. Предложенная метрика IBS_{RM} оценивает среднее интегральное отклонение прогноза функции выживания $S(t|X_i)$ от эталонной функции. Пусть $N(t)$ — количество оставшихся наблюдений к моменту t, t_{max} — правая граница временной шкалы, тогда:

$$IBS_{RM} = \frac{1}{t_{max}} \int_0^{t_{max}} \left(\frac{1}{N(t)} \sum_i \begin{cases} (0 - S(t|X_i))^2, & \text{если } T_i \le t, \delta_i = 1, \\ (1 - S(t|X_i))^2, & \text{если } T_i > t, \\ 0, & \text{если } T_i = t, \delta_i = 0, \end{cases} \right) dt. \quad (4)$$

Предложенная метрика качества $IAUC_{WW,TI}$ оценивает качество ранжирования наблюдений относительно значений прогнозируемого кумулятивного риска $\hat{h}(t|X_i)$ для всех моментов времени. Пусть t_{max} и t_{min} — правая и левая граница временной шкалы, тогда метрика имеет вид:

$$IAUC_{WW,TI} = \int_{t_{min}}^{t_{max}} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} I(T_j > t) I((T_i \le t) \cdot \delta_i) I(\hat{h}(t|X_j) \le \hat{h}(t|X_i))}{(t_{max} - t_{min})(\sum_{j=1}^{n} I(T_j > t))(\sum_{i=1}^{n} I((T_i \le t) \cdot \delta_i))} dt. \quad (5)$$

По результатам исследования, для каждой прогнозируемой величины выделены наиболее стабильные метрики качества: CI для оценки ожидаемого времени события, IBS_{RM} и $AUPRC$ для оценки функции выживания, $IAUC_{WW,TI}$ для оценки функции риска.

В подразделе 3.2 проводится экспериментальное исследование качества прогнозирования методов построения деревьев выживания на метриках CI, IBS_{RM}, $AUPRC$, $IAUC_{WW,TI}$. Постановка эксперимента состоит из трех этапов: выделение обучающей выборки, поиск гиперпараметров по кросс-валидации на обучающей выборке, валидация модели с лучшими гиперпараметрами по 20-кратному семплированию (обучение на тренировочных данных и валидация на тестовой выборке). Использование взвешенных критериев разбиения приводит к улучшению качества на 5 из 6 наборах данных по совокупности метрик. Предложенный метод регуляризации (на этапе поиска разбиения и построения непараметрической модели) позволил улучшить качество предложенной модели дерева выживания и превзойти существующий метод построения деревьев выживания на всех наборах данных.

Четвертая глава посвящена исследованию и разработке методов ансамблирования деревьев выживаемости. Основная задача базовой модели состоит в точном описании соответствующей обучающей подвыборки. Агрегация прогнозов нескольких базовых моделей позволяет улучшить качество прогнозирования и предотвратить переобучение. В отличие от деревьев выживания, ансамблевые модели не имеют интерпретации зависимостей и нацелены на построение точных прогнозов, хотя и позволяют выявить наиболее важные признаки.

В подразделе 4.1 предложен метод построения бутстреп ансамбля независимых деревьев выживаемости. Метод основан на идее построения множества деревьев выживаемости на бутстреп-выборках (с возвращением) и последующим усреднением прогнозов базовых моделей, причем функции выживания и риска усредняются для всех моментов времени. Функция потерь используется для определения размера ансамбля на OOB (out-of-back) выборке. Параллельная реализация алгоритма основана на построении заданного количества деревьев выживаемости с последующим определением размера по глобальному максимуму качества на OOB выборке.

В подразделе 4.2 предложен метод построения адаптивного бустинг ансамбля деревьев выживаемости с перевыборкой. Метод основан на идее итеративного построения ансамбля моделей, в котором каждая последующая модель строится по выборке с наблюдениями, имеющими низкое качество прогноза на предыдущих итерациях ансамбля. Формально, для каждого наблюдения обучающей выборки сопоставляется вес w_i, который определяет вероятность попадания наблюдения в следующую обучающую подвыборку и корректируется на основе ошибки базовой модели. На полученной подвыборке на шаге t строится дерево выживания h_t и вычисляется прогноз $y^{(t)}(X_i)$ для каждого наблюдения. На основе отклонения $L(T_i, \delta_i, y_i^{(t)}(X_i))$ прогноза от истинного значения вычисляется ошибка L_i (при условии $D = \sup_i L(T_i, \delta_i, y_i^{(t)}(X_i))$):

$$L_i^{linear} = \frac{L(T_i, \delta_i, y_i^{(t)}(X_i))}{D},$$

$$L_i^{exp} = 1 - \exp\left[\frac{-L(T_i, \delta_i, y_i^{(t)}(X_i))}{D}\right],$$

$$L_i^{sigmoid} = 1/\left(1 + \exp\left[\frac{-L(T_i, \delta_i, y_i^{(t)}(X_i))}{D}\right]\right).$$

Далее, вычисляется средняя ошибка модели $L = \sum_{i=1}^{N} L_i p_i$ и мера уверенности в прогнозе $\beta_t = L/(1 - L)$, а обновление весов наблюдений обучающей выборки выполняется согласно правилу:

$$w_i \qquad w_i \beta_t^{(1-L_i)}.$$

Таким образом, для наблюдений с меньшей ошибкой прогнозирования уменьшается вероятность попадания в следующую обучающую выборку. Прогнозом для наблюдения с вектором признаков X будет служить взвешенная сумма прогнозов базовых моделей:

$$y(X) = \sum_{t=1}^{M} \frac{log(1/\beta_j)}{\sum_k log(1/\beta_k)} y^{(t)}(X).$$

Необходимым требованием для функции потерь L является возможность расчета индивидульных потерь по единичному наблюдению. Также на качество прогнозирования влияет стратегия обновления весов. При глобальной стратегии формула обновления весов применяется для всех наблюдений обучающей выборки, а для локальной стратегии только для наблюдений, попавших в последнюю подвыборку. Экспериментально продемонстрировано, что глобальная стратегия приводит к появлению редких наблюдений с высокой ошибкой. Данные наблюдения имеют высокую вероятность попадания в обучающую подвыборку и приводят к построению схожих базовых моделей.

В подразделе 4.3 проводится экспериментальное исследование качества прогнозирования методов анализа выживаемости. Постановка эксперимента и используемые метрики качества повторяют подраздел 3.2. Первой целью экспериментального исследования является анализ влияния модификаций существующих метрик на качество бутстреп ансамбля с последующим выбором лучшей функции потерь L. По результатам экспериментального исследования, среди функций потерь наибольшее качество достигается на предложенной метрике IBS_{RM}. Функция потерь используется для определения размера ансамбля, а также для расчета ошибки L_i в модели адаптивного бустинга с перевыборкой.

Второй целью экспериментального исследования является оценка и сравнение качества существующих и предложенных моделей анализа выживае-

Таблица 1: Сравнение средних рангов качества по всем наборам данных

Модель	CI	$\boldsymbol{IBS_{RM}}$	$\boldsymbol{IAUC_{WW,TI}}$	$\boldsymbol{AUPRC}$
$KaplanMeier$	15.0	15.5	16.0	15.5
$CoxPH$	9.5	10.5	9.67	10.17
$LogLogisticAFT$	9.17	8.67	8.0	11.33
$LogNormalAFT$	10.67	11.17	10.33	11.33
$WeibullAFT$	9.67	9.5	8.33	11.67
ST	14.5	14.0	14.33	12.67
RSF	8.33	11.67	7.67	12.33
$CWGBSA$	10.5	11.67	10.83	12.83
$GBSA$	7.33	8.83	8.0	9.83
$DeepSurv$	10.17	8.5	8.33	6.17
$CoxTime$	9.17	9.17	8.67	6.17
$TREE_{KMWV}$	7.83	6.83	9.67	**1.0**
$Bootstrap$	**3.67**	**2.67**	**4.33**	**2.83**
$Boosting_{linear}$	3.83	**1.83**	**3.67**	5.5
$Boosting_{exp}$	**3.33**	**2.17**	**3.67**	4.5
$Boosting_{sigmoid}$	**3.33**	3.33	4.5	**2.17**

мости. Рассматриваются существующие статистические модели: непараметрические (KaplanMeier), полупараметрические (CoxPH), параметрические (LogLogisticAFT, LogNormalAFT, WeibullAFT), и модели машинного обучения: деревья выживаемости (ST), ансамбли (RSF, CWGBSA, GBSA) и нейронные сети (DeepSurv, CoxTime). Для каждого набора данных были определены порядки (ранги) ранжирования моделей по качеству прогнозирования по каждой метрике. В таблице 1 представлены значения рангов моделей, усредненных по всем наборам данных. Лучшее качество определяется наименьшим рангом, а жирным отмечены лучшие 3 значения. Предложенная модель дерева выживаемости $TREE_{KMWV}$ превзошла по качеству существующую модель дерева выживаемости ST. Предложенные модели ансамблирования превзошли по среднему рангу существующие методы. Модель бутстреп ансамбля $Bootstrap$ демонстрирует лучшее качество по совокупности метрик, в то время как модель адаптивного бустинга с перевыборкой $Boosting$ позволяет повысить качество по метрикам при различных весовых схемах.

Пятая глава посвящена разработке и реализации открытой библиотеки анализа выживаемости survivors, использующей предложенный комплекс

алгоритмов. Приводится детальное описание архитектуры, программной реализации разработанной библиотеки и сценариев использования.

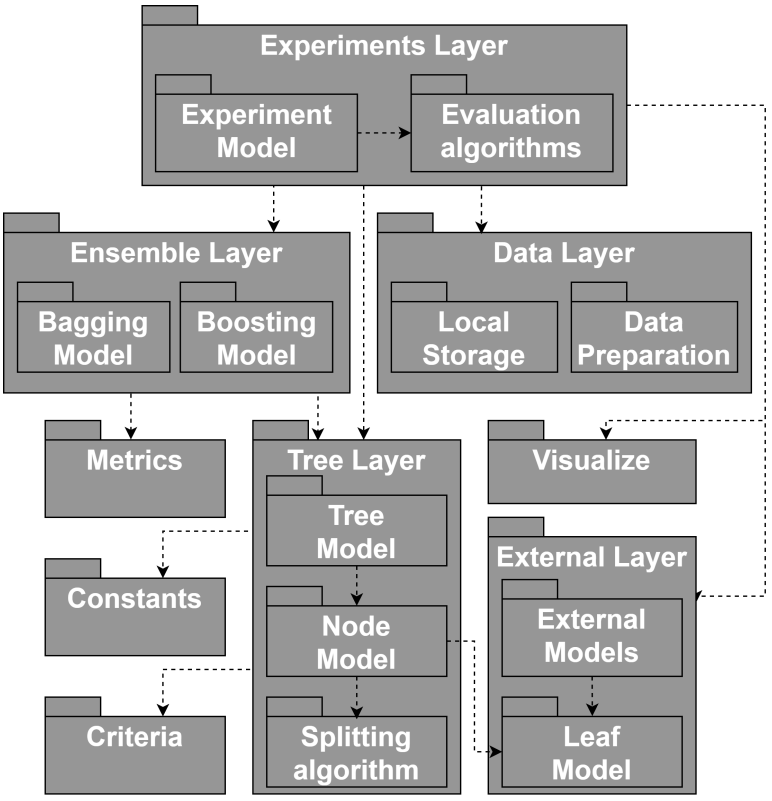


Рис. 3: Диаграмма пакетов библиотеки `survivors` на универсальном языке моделирования (Unified Modeling Language, UML).

В подразделе 5.1 рассматриваются достоинства и недостатки существующих библиотек анализа выживаемости с открытым исходным кодом: Scikit-survival, Lifelines, PyCox. Подраздел 5.2 содержит описание архитектуры библиотеки анализа выживаемости и программных компонентов. Предлагаемое архитектурное решение состоит из следующих модулей (Рисунок 3):

1. Модуль внутреннего представления (Constants) содержит константы и функции, необходимые для работы с внутренним представлением данных и прогнозными моделями.
2. Модуль метрик качества (Metrics) содержит реализации классических метрик анализа выживаемости, а также их модификации, разработанные в третьей главе при исследовании случаев избыточной чувствительности метрик.
3. Модуль статистических критериев (Criteria) содержит векторизованные реализации статистических критериев: log-rank, wilcoxon, tarone-ware, peto-peto и др. Ускорение вычислений осуществляется за счет использования JIT-компиляции (Just-In-time) исходного кода в байт-код.

4. Модуль визуализации (Visualize) содержит функции вывода результатов работы алгоритмов в удобном формате для пользователя. Выделяются следующие группы функций: визуализация прогнозов, анализ поведения метрик качества, визуализация характеристик данных.
5. Модуль внешних моделей (External Layer) содержит реализации существующих моделей анализа выживаемости с унифицированным интерфейсом LeafModel.
6. Модуль построения дерева выживаемости (Tree Layer) содержит описание основных сущностей и реализацию предложенных методов выбора лучшего разбиения выборки и построения деревьев выживания. Класс Node отвечает за отдельные структурные элементы дерева выживаемости, а также за поиск правила разбиения узловой выборки. Класс CRAID определяет структуру дерева выживаемости.
7. Модуль построения ансамблей деревьев решений (Ensemble Layer) содержит базовый класс BaseEnsemble, а также реализацию интерфейса в виде модели бутстеп ансамбля BootstrapCRAID и модели адаптивного бустинга с перевыборкой BoostingCRAID.
8. Модуль сбора наборов данных (Data Layer) предназначен для загрузки и предобработки открытых медицинских наборов данных. На этапе предобработки исходные данные приводятся к унифицированной структуре: X (пространство признаков), y (упорядоченный массив с двумя целевыми переменными), features (исходные названия признаков наблюдений), categ (подмножество категориальных признаков).
9. Модуль экспериментов (Experiments Layer) содержит класс Experiments для запуска экспериментов с различными стратегиями поиска лучших гиперпараметров, валидации и сравнения моделей выживаемости.

В подразделе 5.3 описываются базовые сценарии функционирования библиотеки анализа выживаемости: сбор и подготовка данных, построение статистических моделей, построение деревьев выживаемости и интерпретация зависимостей, построение ансамблей деревьев выживаемости, оценка качества прогнозирования.

В подразделе 5.4 проводится экспериментальная оценка производительности предложенной программной реализации. Для предложенных методов построения дерева выживаемости, бутстреп ансамбля и адаптивного бустинга с перевыборкой вычисляется время работы и объем потребляемой памяти при обучении и прогнозировании. Полученные оценки свидетельствуют о возможности применения разработанной библиотеки на практике.

В заключении формулируются основные результаты работы.

Основные результаты работы

1. Предложен метод построения деревьев выживаемости, основанный на алгоритме поиска лучшего разбиения с учетом взвешенных регуляризованных log-rank критериев. Взвешенные критерии позволяют придавать разный приоритет ранним и поздним событиям, повышая чувствительность к распределению вероятностей времени наступления событий. Регуляризация критерия разбиения позволяет учитывать информацию об априорном распределении времени наступления событий. Метод способен обрабатывать категориальные признаки и пропуски в данных, а также применим к случаям информативного цензурирования.
2. Предложены методы ансамблирования деревьев выживаемости. Бутстреп метод основан на построении ансамбля независимых деревьев на бустреп-выборках с выбором размера по минимальной ошибке вне бустреп-выборки. Модель бустинга основана на построении адаптивного ансамбля деревьев с перевыборкой, в котором каждая последующая модель обучается на наиболее сложных для прогнозирования ансамблем наблюдениях. Выбор функции потерь проводился на основе исследования чувствительности метрик качества к особенностям данных.
3. Реализована открытая программная библиотека анализа выживаемости, использующая предложенный комплекс алгоритмов. Библиотека позволяет проводить сценарии сбора и предобработка данных, построения и применения моделей выживаемости, оценки качества прогнозирования. По результатам проведенных на основе библиотеки экспериментальных исследований на медицинских данных, разработанные методы превзошли по качеству прогнозирования существующие методы анализа выживаемости.

Публикации по теме диссертации

Научные статьи, опубликованные в рецензируемых журналах, индексируемых в международных базах Scopus, WoS, RSCI

1. Vasilev I., Petrovskiy M., Mashechkin I. Sensitivity of survival analysis metrics // Mathematics. – 2023. – Vol. 11, no. 20. – P. 4246. – (WoS Q1: JIF – 2,3; 2,125/2,0).

   Автором были самостоятельно проанализированы случаи избыточной

чувствительности метрик качества анализа выживаемости и разработаны модифицированные метрики с равным вкладом событий и временных интервалов. Задача была поставлена и сформулирована совместно с М. И. Петровским и И. В. Машечкиным.

2. Vasilev I., Petrovskiy M., Mashechkin I. Adaptive sampling for weighted log-rank survival trees boosting // Lecture Notes in Computer Science. – 2023. – Vol. 13822. – P. 98-115 – (Scopus: SJR – 0,61; 1,125/1,0).
   Автором был самостоятельно разработан метод адаптивного бустинга деревьев выживаемости с перевыборкой. Анализ и интерпретация результатов экспериментов проводилась совместно с М. И. Петровским и И. В. Машечкиным.
3. Васильев Ю. А. Разработка библиотеки древовидных моделей анализа выживаемости // Вестник Московского университета. Серия 15: Вычислительная математика и кибернетика. – 2024. – № 3. – С. 60-72 – (RSCI: импакт-фактор РИНЦ – 0,142; 0,8125/0,8125).
   Перевод:
   Vasilev I. A. Developing a Library of Tree-Based Models for Survival Analysis // Moscow University Computational Mathematics and Cybernetics. – 2024. – Vol. 48, no. 3. – P. 190-202 – (ВАК: импакт-фактор отсутствует; 0,8125/0,8125).
   Автором была самостоятельно разработана программная библиотека древовидных моделей анализа выживаемости, а также предложен гистограммный метод поиска разбиения выборок с цензурированием.
4. Васильев Ю.А., Петровский М.И., Машечкин И.В. Применение регуляризации при вычислении критериев разбиения в моделях анализа выживаемости // Вычислительные методы и программирование. – 2024. – Т. 25, № 3. – С. 357-377 – (RSCI: импакт-фактор РИНЦ – 0,511; 1,3125/1,2).
   Автором был разработан подход регуляризации критерия log-rank для построения деревьев выживаемости. Анализ и интерпретация случаев информативного цензурирования были выполнены совместно с М. И. Петровским и И. В. Машечкиным. Программная реализация и численные эксперименты выполнены автором полностью самостоятельно.

Иные публикации

Свидетельства о регистрации программных комплексов:

1. Свидетельство о государственной регистрации программы для ЭВМ No.2024681935 «Библиотека методов машинного обучения для построения моделей анализа выживаемости». Правообладатель: Васильев

Юлий Алексеевич. Заявка No. 2024680633. Дата государственной регистрации в Реестре программ для ЭВМ 16 сентября 2024г.