ОТЗЫВ ОФИЦИАЛЬНОГО ОППОНЕНТА

на диссертационную работу Васильева Юлия Алексеевича «Исследование и разработка методов машинного обучения анализа выживаемости», представленную на соискание ученой степени кандидата физико-математических наук по специальности 2.3.5 «Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей»

Интеллектуальные и статистические методы анализа событий играют критически важную роль в современном мире и позволяют описывать контекст события, интерпретировать зависимости и прогнозировать наступление события на основе характеристик наблюдения. Такие методы особенно часто применяются в здравоохранении, теории надежности, маркетинге и социологии.

Одним из ключевых направлений в этой области является анализ выживаемости. Задачей методов анализа выживаемости является оценка вероятности и времени до наступления определенного случайного события. Важно отметить, что такие методы работают с цензурированными наблюдениями, для которых истинное время наступления события неизвестно из-за выхода наблюдения из исследования или окончания исследования.

Также методы анализа выживаемости позволяют оценивать изменение вероятности наступления события с течением времени, прогнозируя функции выживания и риска. Для применения моделей выживаемости на практике необходимо обеспечить прогноз значений функций выживания или риска для каждого момента времени, поскольку использование дискретной временной шкалы может быть недостаточно для оценки темпов изменения рисков наступления события.

Применение методов анализа выживаемости часто связано с обработкой реальных данных. Например, сведения о пациенте могут быть заполнены не полностью, содержать количественные и категориальные значения. Классические методы анализа выживаемости работают только с числовыми признаками наблюдения, и для применения данных методов к реальным данным требуется предварительная обработка сырых данных, которая может приводить к потере или искажению информации.

Другой проблемой классических подходов является использование статистических предположений, которые могут быть не применимы к реальному распределению вероятностей времени наступления события или цензурирования наблюдений. К таким предположениям относится пропорциональность рисков, неинформативность цензурирования и фиксированное модельное распределение вероятностей времени наступления событий.

При обосновании актуальности темы автор диссертации справедливо обращает внимание на необходимость разработки новых интеллектуальных методов

анализа выживаемости, которые могли бы работать с реальными данными без предварительной обработки и без использования ограничивающих предположений.

Таким образом, тема диссертационной работы Васильева Ю.А., связанная с разработкой математического и программного обеспечения для решения задач анализа выживаемости на реальных данных, является безусловно актуальной.

Диссертационная работа состоит из введения, пяти глав, заключения и списка литературы. Полный объём диссертации насчитывает 142 страницы. Список литературы состоит из 123 наименований. Во введении формулируются цели диссертационной работы, обосновывается актуальность работы, перечисляются рассматриваемые задачи и полученные результаты.

В первой главе проводится впечатляющий аналитический обзор методов сбора событийных данных и анализа выживаемости. Рассматривается широкий спектр методов построения статистических моделей и моделей машинного обучения непрерывного и дискретного времени, описываются их преимущества и недостатки. В частности, подробно рассматриваются недостатки статистических предположений пропорциональных рисков, критерия разбиения log-rank и теоретической формы (модельного) распределения вероятности времени, а также их роль в методах машинного обучения. Отдельное внимание уделяется обзору метрик анализа выживаемости, даются рекомендации об использовании интегральных и точечных метрик для оценки качества прогнозирования.

Вторая глава посвящена разработке методов построения деревьев выживаемости, применимых к особенностям реальных данных. Подтверждается наличие особенностей в шести различных медицинских наборах данных. Предлагается метод поиска лучшего правила разбиения данных с цензурированием по категориальным и непрерывным признакам со сравнением функций риска по взвешенному критерию log-rank. Предлагается метод построения деревьев выживаемости на основе рекурсивного поиска лучшего правила разбиения выборки среди всех признаков. Отдельно рассматриваются ограничения критерия log-rank и непараметрической модели Каплана-Мейера при работе с данными с информативным цензурированием и предлагаются способы их преодоления. В результаты предлагается метод построения деревьев выживаемости возможно без предобработки данных и использования статистических предположений.

Третья глава посвящена исследованию и разработке методов оценки качества прогнозирования моделей анализа выживаемости. В главе выделены и подробно рассмотрены четыре случая избыточной чувствительности метрик качества. Доказано наличие такой чувствительности у рассматриваемых метрик, в том числе — на примере реальных данных. Для каждого случая предлагаются модификации метрик качества для уравнивания вклада отдельных событий при валидации. Выделяются четыре метрики качества, обладающие наибольшей устойчивостью к особенностям реальных данных. Также проводится экспериментальное сравнение методов построения деревьев выживаемости, в результате которого предложенный в диссертации метод достигает наилучшего качества прогнозирования.

Четвертая глава посвящена разработке ансамблей деревьев выживаемости. Предлагаются два метода ансамблирования путем агрегации прогнозов независимых моделей и построения адаптивного ансамбля с корректировкой вероятностей попадания наблюдений в следующую обучающую выборку. Большая часть главы отводится на экспериментальное исследование влияния функции потерь в ансамблях на качество прогнозирования и сравнение с существующими статистическими подходами и методами машинного обучения. Предложенные в диссертации ансамбли превзошли существующие модели по качеству прогнозирования на всех наборах данных.

В пятой главе описывается разработанная программная библиотека анализа выживаемости, ее архитектура и модули. Библиотека содержит комплекс предложенных алгоритмов и метрик и имеет открытый исходный код. Проводится экспериментальная оценка производительности предложенных моделей.

В заключении сформулированы в трех пунктах основные результаты работы. Заключение полностью согласуется с содержанием диссертации.

К содержанию оппонируемой диссертационной работы у меня есть несколько замечаний:

1. На стр.16 символом «дельта» обозначен флаг цензурирования, а на стр.20 — флаг наступления события, что противоречит друг другу.
2. На стр.19 неточность формулировке во фразе «...склонны завышать функцию выживания (приближая к константной 1)». Видимо, под «константной» автор имел ввиду «постоянную» или «константу», но функция выживания не может быть всюду постоянной. Также термин «константный» вместо «постоянный» встречается на стр.34, 67, 76.
3. На стр. 20 в формуле для логарифма функции правдоподобия пропущены крышечки над h, S, обозначающие использование соответствующих статистических оценок.
4. И далее в разделе 1.3. выбраны неудачные обозначения, которые могут ввести в заблуждение, потому что меняются от формулы к формуле: непараметрические оценки функции выживания и риска ранее обозначались через h и S с крышечками, в разделе 1.3.4. вместо крышечек использован индекс «0» ($S_0(t)$ и $h_0(t)$), а в разделе 1.3.5 эти же символы $S_0(t)$ и $h_0(t)$ используются уже для обозначения модельных функций выживания и риска.
5. На стр. 25 при описании метода построения оценки функции выживания по таблицам времен жизни содержится фраза: «Поскольку вероятности выживания считаются независимыми на разных интервалах, доля равна произведению долей выживших объектов по всем предыдущим интервалам». Здесь некорректное употребление термина «независимость» применительно к вероятностям (числам). Вероятно, автор имел ввиду независимость событий, состоящих в наступлении терминального события на разных интервалах. Но эти события не являются независимыми (например, потому что если терминальное событие наступило на каком-то

интервале, то оно не может наступить на последующих). Вычисление функции выживания в виде произведения долей основано на теореме умножения теории вероятностей, согласно которой вероятность произведения событий (выживания на каждом из рассматриваемых интервалов) равна произведению *условных* вероятностей (долей на каждом интервале) — именно они и содержатся в таблице.

6. На стр.26 медиана (корень уравнения $S(t)=0.5$) названа ожидаемым времени жизни, т. е. математическим ожиданием (а оно определяется как интеграл, а не корень уравнения на значение функции выживания).
7. При описании таблицы 3 на стр.47 имеется двусмысленность в количестве признаков в исследовании WUHAN: в таблице указано 224 признака, тогда как в тексте упоминается, что признаков было всего 76.
8. На рис. 9 стр.49 приведены графики ядерных оценок плотности распределения моментов времени наступления событий и цензурирования, «вылезающие» на отрицательную полуось. Происхождение этого артефакта, по-видимому, объясняется сглаживанием за счет использования ядерной оценки плотности и не имеет отношения к реальности. Этот эффект не мешало бы пояснить в тексте диссертации, чтобы не смущать читателя отрицательными временами жизни.
9. При описании предложенной модификации непараметрической оценки Каплана-Мейера (раздел 2.4.4.) использования нормального закона для симуляции времени виртуальных событий не обосновано и, более того, не вполне адекватно, так как по смыслу сгенерированные случайные величины должны быть неотрицательными, а нормальное распределение сосредоточено на всей прямой. Более естественным, на мой взгляд, было использование распределений с неотрицательным носителем, например, гамма или Вейбулла. Так же интересно было бы попробовать использовать эмпирическое распределение, построенное по доступной выборке.

Перечисленные замечания не умаляют научной ценности диссертационного исследования и не влияют высокую научную оценку полученных результатов.

В заключение отмечу, что диссертация представляет собой завершенное научное исследование, обладающее внутренним единством. Результаты диссертационной работы имеют не только научную, но и практическую значимость и новизну и могут стать основой для создания современных систем анализа событий, включающих средства анализа выживаемости, например, в медицине, теории надежности, маркетинге и других областях. Все научные положения, выводы и рекомендации, сформулированные в диссертации, полностью обоснованы, достоверны и подтверждены экспериментами на реальных данных. Основные результаты работы опубликованы в четырех статьях в рецензируемых изданиях. Также автором зарегистрированы права на программу для ЭВМ. Автореферат правильно отражает содержание диссертации.

Диссертация Васильева Юлия Алексеевича «Исследование и разработка методов машинного обучения анализа выживаемости» удовлетворяет пп. 2.1-2.5

Положения о присуждении ученых степеней в Московском государственном университете имени М.В. Ломоносова и соответствует специальности 2.3.5 «Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей». Диссертация оформлена согласно требованиям Положения о совете по защите диссертаций на соискание ученой степени кандидата наук Московского государственного университета имени М.В. Ломоносова.

Таким образом, считаю, что представленная диссертационная работа удовлетворяет всем требованиям, установленным Московским государственным университетом имени М.В. Ломоносова к кандидатским диссертациям, а ее автор Васильев Юлий Алексеевич несомненно заслуживает присуждения ученой степени кандидата физико-математических наук по специальности 2.3.5 «Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей».

Официальный оппонент:
Доктор физико-математических наук,
профессор кафедры математической статистики факультета вычислительной математики и кибернетики МГУ имени М.В. Ломоносова

Шевцова Ирина Геннадьевна

13 декабря 2024 г.

Контактные данные:
Тел.: +7(495)939-53-94, email: ishevtsova@cs.msu.ru
Специальность, по которой официальным оппонентом защищена диссертация:
01.01.05 «Теория вероятностей и математическая статистика»

Адрес места работы:
Москва, 119991, Ленинские горы, МГУ имени М.В. Ломоносова, 2-й учебный корпус, факультет ВМК, кафедра математической статистики
Тел.: +7(495)939-53-94, email: ishevtsova@cs.msu.ru

Подпись профессора кафедры математической статистики факультета вычислительной математики и кибернетики МГУ имени М.В. Ломоносова Шевцовой Ирины Геннадьевны заверяю

ПОДП
Зав.
И.С.
Подпись