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»

Диссертационная работа Ю.А. Васильева посвящена решению задачи анализа событий на основе методов машинного обучения анализа выживаемости. В такой постановке возможно оценка не только вероятности и времени до наступления события, но и функций выживания и риска, отражающих динамику вероятности события в течение времени. В отличии от методов классификации и регрессии, методы анализа выживаемости используют две целевые переменные: время наступления события с момента входа в исследование и индикатор наступления события. Также, в отличии от методов анализа временных рядов, модели выживаемости используют информацию только при входе наблюдения в исследование, что упрощает сбор данных и применение моделей. Особенно актуально применение данных моделей для анализа медицинских данных пациентов, оценки эффективности схем лечения и поддержки врачебных решений.

Автор отмечает, что применение существующих методов к реальным данным осложнено из-за структуры данных и строгих статистических предположений. Следовательно, основное внимание в работе уделяется исследованию и разработке древовидных подходов, которые не имеют априорных статистических предположений и применимы к сырым данным. Таким образом, используются и преимущества данных подходов, поскольку деревья решений позволяют строить интерпретируемые прогнозы, а ансамбли деревьев решений обладают высокой точностью прогнозирования. Высокая актуальность направления выполненных исследований сомнений не вызывает.

Обратимся к более подробному изложению содержания диссертации по главам.

В первой главе проводится обзор существующих методов анализа выживаемости, включая статистические подходы и специализированные методы машинного обучения. При описании методов особое внимание уделяется их недостаткам: статистические предположения часто не выполняются на реальных данных, а методы машинного обучения либо наследуют статистические предположения, либо сводят задачи к классификации или регрессии на дискретных временных шкалах. Помимо этого, отдельно рассматриваются оценки качества прогнозирования различных величин анализа выживаемости, подчеркивается необходимость исследования чувствительности метрик к особенностям данных.

Вторая глава посвящена общей идее разработке метода построения деревьев выживаемости. Для обработки гетерогенных данных предлагается метод поиска лучшего разбиения выборки, применимый к непрерывным и категориальным признакам с возможными пропусками. В листьях дерева выживаемости строятся непараметрические оценки функций выживания и риска. Для обработки случаев мультимодального распределения вероятностей наступления времени событий предлагается подход регуляризации критерия разбиения, а также модификация непараметрической модели.

В третьей главе поднимается вопрос разработки методов оценки качества прогнозирования величин анализа выживаемости. Автор отмечает, что существующие метрики определяют высокую значимость редких поздних событий и временных интервалов и не могут быть использованы для валидации моделей. Предложенные модификации определяют равный вклад наблюдений при расчете метрик. Далее проводится экспериментальное исследование качества предложенных методов и модификаций, а также сравнение с существующим подходом построения дерева выживаемости. Разработанные автором методы позволяют повысить качество прогнозирования по четырем метрикам, устойчивых к особенностям реальных данных, на шести медицинских наборах данных.

Четвертая глава посвящена разработке ансамблей предложенных деревьев выживаемости. Первый подход основан на классическом усреднении прогнозов базовых моделей, построенных на бутстеп-выборках, и выбором лучшего количества деревьев по качеству прогнозирования ансамбля. Второй

подход использует эмпирическую схему итеративного обновления весов наблюдений для построения адаптивного ансамбля с перевыборкой. В рамках данной главы также поднимается вопрос выбора функций потерь при обучении ансамбля, причем лучшее качество прогнозирования достигается при использовании предложенной метрики. Также, предложенные ансамбли превзошли существующие методы.

Наконец, в пятой главе представлен разработанный программный комплекс в виде библиотеки с открытым кодом. Демонстрируется то, как возможности языка Python позволяют использовать функционал библиотеки в рамках рассматриваемого класса прикладных задач. В главе также проводится оценка времени выполнения и используемой памяти для предложенных алгоритмов построения деревьев выживаемости и их ансамблей.

Достоверность и обоснованность полученных в работе результатов обеспечена согласованностью построенных моделей и используемых данных, подробному описанию серии проведенных вычислительных экспериментов на шести реальных медицинских наборах данных, а также соответствию результатов экспериментов итоговым выводам.

Новизна данной работы заключается в разработке новых методов построения деревьев выживаемости и их ансамблей для работы с реальными данными. В частности, новыми являются предложенные методы поиска лучшего разбиения неполных разнородных данных с цензурированием, регуляризации критерия разбиения дерева выживаемости, модификации непараметрических оценок функции выживания и риска для обработки случаев информативного цензурирования, а также метод построения адаптивного ансамбля деревьев выживаемости с перевыборкой. Результаты исследования чувствительности существующих метрик качества анализа выживаемости и разработанные модификации метрик также являются новыми.

Научная и практическая значимость работы состоит в том, что предложенные в рамках работы методы и подходы реализованы в виде открытой программной библиотеки, могут быть использованы для решения актуальных задач анализа выживаемости, при этом методы имеют допустимую производительность для их применения на практике.

Основные результаты опубликованы в 4 публикациях, из которых 4 – в изданиях WoS (одна статья опубликована в журнале Q1), Scopus, RSCI, рекомендованных для защиты в диссертационном совете МГУ им. М.В. Ломоносова по специальности 2.3.5. Также, зарегистрированы права на программное обеспечение для ЭВМ. Автореферат в полной мере отражает содержание диссертационной работы.

Замечания по диссертационной работе:

1. В третьей главе проводится исследование чувствительности метрик анализа выживаемости к распределению вероятностей времени наступления событий, временным интервалам и типу событий. Однако, в ходе исследования не поднимается вопрос о связи существующих и модифицированных метрик между собой. Было бы полезно более детально обсудить этот аспект при выборе метрик качества и функций потерь для экспериментального исследования качества прогнозирования и сравнения моделей.

2. В работе рассматриваются шесть медицинских наборов данных относительно небольшого размера (не более десяти тысяч наблюдений). Для демонстрации преимуществ предложенных методов было бы полезно провести экспериментальное исследование на более крупных наборах данных.

3. При анализе существующих статистических моделей автор отмечает, что модель должна обеспечивать возможность пересечения прогнозов функций выживания для различных наблюдений, однако аргументация данного требования не описывается в работе напрямую. В разделе 1.3.4, на странице 30, автор отмечает, что «в клинической практике, пересекаемость функций выживания двух схем лечения говорит об отсутствии различий между схемами». Следует дополнительно обосновать необходимость данного требования при выделении недостатков существующих моделей.

Вместе с тем, указанные замечания к диссертации не умаляют значимости диссертационного исследования. Диссертация отвечает требованиям, установленным Московским государственным университетом имени М.В. Ломоносова к работам подобного рода. Содержание диссертации соответствует специальности 2.3.5 – «Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей» (по физико-математическим наукам), а также критериям, определенным пп. 2.1-2.5

Положения о присуждении ученых степеней в Московском государственном университете имени М.В. Ломоносова, а также оформлена согласно требованиям Положения о совете по защите диссертаций на соискание ученой степени кандидата наук Московского государственного университета имени М.В. Ломоносова.

Таким образом, соискатель Васильев Юлий Алексеевич 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, доцент
Академический руководитель направления [illegible] Акционерного общества "ТБанк"

Дьяконов Александр Геннадьевич

0 декабря 2024 г.

Контактные данные:
Тел.: +7 916 160 17 02, email: a.g.dyakonov@tbank.ru
Специальность, по которой официальным оппонентом защищена диссертация:
01.01.09 – «Дискретная математика и математическая кибернетика»

Адрес места работы:
127287, Москва, Ул. 2-я Хуторская, д.38А, стр.26,
Акционерное общество "ТБанк"
Тел.: +7 916 160 17 02, email: a.g.dyakonov@tbank.ru