РАЗРАБОТКА АЛГОРИТМА ПРЕДСКАЗАНИЯ ВЫЖИВАЕМОСТИ ПАЦИЕНТОВ С ОНКОЛОГИЧЕСКИМИ ЗАБОЛЕВАНИЯМИ

В. Н. Яцков, М. К. Чепелева

Белорусский государственный университет, Минск, Беларусь
E-mail: vlad18742@gmail.com, maryna.chepeleva@gmail.com

Разработан алгоритм для предсказания выживаемости пациентов с онкологическими заболеваниями методом блочного леса с расщеплением узлов по методу exponential log-likelihood loss. Проведено сравнение качества предсказания с алгоритмами случайного леса выживаемости и бустинга регрессии кокса. "Получена лучшая точность 84,04 % по Бриеру и 98,89 % по С-индексу для блочного леса". Разработанные программные средства могут быть использованы для предсказания клинических рисков в персонализированной медицине.

Ключевые слова: *случайный лес выживаемости; блочный лес; секвенирование; предсказание выживаемости.*

ВВЕДЕНИЕ

С развитием биомедицинских технологий все большее распространение получает подход персонализированной медицины (*personalized medicine*), который предполагает индивидуальное рассмотрение многомерных данных конкретного пациента для принятия решений [1]. Такой подход становится особенно актуальным при онкологических заболеваниях ввиду высокой гетерогенности опухолей, и одной из его составляющих является предсказание рисков для пациентов.

Предсказание выживаемости подразумевает оценку времени наступления критического события, основанную на функции вероятности наступления события. Прогнозирование функции выживаемости пациента и установление влияния признаков (в том числе видов терапии) позволяют принять решение об оптимальном плане лечения.

Метод случайного леса отличается относительно простой установкой связи между ковариатами и риском, а также показывает высокую точность прогнозирования [2].

Цель данной работы – разработка алгоритма для предсказания выживаемости пациентов, больных раком молочной железы, на основе метода блочного леса, сравнение точности предсказания со случайным лесом выживаемости и бустингом регрессии кокса.

МАТЕРИАЛЫ И МЕТОДЫ

Для тестирования алгоритмов были использованы данные секвенирования РНК для пациентов с раком груди [3]. После очистки от неинформативных данных выборка состояла из 1158 образцов. Для 198 образцов наступило критическое событие в определенный момент времени, 960 – цензурированы. Критическим событием является смерть пациента. Помимо экспрессии генов имелся набор клинических признаков: пол, подтип рака, тип образца ткани, группа по наличию раковой опухоли.

Случайный лес выживаемости и блочный лес

Расширение метода случайного леса Брэймана для цензурированных справа данных событийно-времязависимой (*time-to-event*) информации строится на основе рекурсивного разделения ковариантного пространства для формирования групп субъектов, похожих по *time-to-event* результату [4].

Блочный лес – модификация случайного леса выживаемости, в котором для повышения точности работы исходные признаки случайно распределяются по блокам. Каждому блоку присваивается весовой коэффициент $w_m \in (0, 1]$. Все признаки, принадлежащие m-му блоку, также имеют весовой коэффициент w_m, который изменяет значимость выбранного признака при расщеплении узлов деревьев [5].

Алгоритм случайного леса выживаемости был реализован на основе *R*-пакета *rpart,* алгоритм блочного леса – на основе *R*-пакета *BlockForest.* В обоих алгоритмах был запрограммирован метод расщепления узлов *exponential log-likelihood loss*.

Алгоритм расщепления узла

Основываясь на сравнительном обзоре алгоритмов расщепления узлов для случайных деревьев выживаемости [2], для расщепления узлов дерева выбран и реализован алгоритм *exponential log-likelihood loss* (*EL*).

Пусть на основе объектов L строится дерево с конечным количеством узлов H. Тогда на узле $h \in H$ находятся объекты $L_h \in L$. При этом каждый объект $l_i \in L_h$ характеризуется параметрами δ_i (результат наступления события) и t_i (время наступления события).

Тогда оцениваемый риск в узле h определяется как

$$\hat{\lambda}_h = \frac{\sum_{l_i \in L_h} \delta_i}{\sum_{l_i \in L_h} t_i}., \qquad (1)$$

В [2] вводят функцию потерь для рассматриваемого узла:

$$R(h) = \sum_{l_i \in L_h} \delta_i - \sum_{l_i \in L_h} \delta_i \ln\left(\hat{\lambda}_h\right). \quad (2)$$

Потери на узле используются в качестве меры ошибки. Расщепление, которое минимизирует потери, считается наилучшим. Дополнительно, в алгоритме блочного леса производится умножение потерь признаков (2) из m-го блока на весовой коэффициент w_m данного блока.

Качество предсказания

Ошибка предсказания выживаемости оценивалась с помощью *C*-индекса и интегрированной оценки Бриера.

C-индекс (concordance index) оценивает вероятность того, что в i-ой паре объектов выполняется условие $S_{i1}(t_{i1}) < S_{i2}(t_{i2})$ при $t_{i1} < t_{i2}$ [4], где $S(t)$ – вероятность выживания объекта на момент времени t. В результате полностью достоверного предсказания значение ошибки по *C*-индексу равно 0, в результате полностью недостоверного – 1.

Оценка Бриера – мера среднеквадратичного отклонения оценки вероятностной величины от ее истинного значения в заданный момент времени. Интегрированная оценка Бриера (*IBS*) – сумма оценок Бриера за все время наблюдения. В результате полностью достоверного предсказания значение данной оценки равно 0, в противном случае – 1.

РЕЗУЛЬТАТЫ

Выживаемость пациентов предсказывалась тремя алгоритмами: улучшенные *EL* расщеплением узлов случайный лес выживаемости и блочный лес, а также алгоритмом бустинга регрессии Кокса из пакета *mboost*. В таблицах 1 и 2 представлено сравнение распределений ошибок предсказания выживаемости пациентов методами блочного леса (*BlockForest*), бустинга регрессии Кокса (*Cox boost*) и случайного леса (*Rpart*). Для *BlockForest* и *Rpart* было использовано 150 деревьев. Данные распределения были получены путем многократных запусков сравниваемых алгоритмов на одинаковых выборках из набора данных.

Таблица 1

Распределения ошибок по *IBS* для сравниваемых алгоритмов

Алгоритм	Минимум	1-й квартиль	Медиана	Среднее	3-й квартиль	Максимум
BlockForest	0,1248	0,1418	0,1564	0,1595	0,1697	0,2303
Cox boost	0,1282	0,1639	0,1811	0,1778	0,1885	0,2477
Rpart	0,1380	0,1588	0,1700	0,1699	0,1790	0,2222

Таблица 2

Распределения ошибок по *C*-индексу для сравниваемых алгоритмов

Алгоритм	Минимум	1-й квартиль	Медиана	Среднее	3-й квартиль	Максимум
BlockForest	0.007059	0.009321	0.011063	0.011082	0.013355	0.015308
Cox boost	0.1575	0.1745	0.1773	0.1884	0.2012	0.2518
Rpart	0.07151	0.10426	0.12424	0.12010	0.13907	0.15480

Получено, что алгоритм блочного леса с алгоритмом расщепления *EL* имеет наименьшую ошибку предсказания среди трех сравниваемых алгоритмов.

ЗАКЛЮЧЕНИЕ

Было проведено сравнение алгоритмов блочного леса, случайного леса выживаемости и бустинга регрессии Кокса. Алгоритмы блочного и случайного леса были улучшены расщеплением узла *exponential log-likelihood loss*. По полученным функциям выживаемости алгоритм блочного леса имеет наилучшую точность предсказания, в среднем равную 84,04 % по оценке Бриера и 98,89 % по *C*-индексу. Разработанный алгоритм предсказания выживаемости может быть встроен в анализ многомерных данных пациентов с онкологическими заболеваниями для предсказания клинических рисков в персонализированной медицине.

БИБЛИОГРАФИЧЕСКИЕ ССЫЛКИ

1. Ma J., Hobbs B.P., Stingo F. C. Statistical Methods for Establishing Personalized Treatment Rules in Oncology // Biomed Res Int. 2015. Vol. 2015, №670691. DOI:10.1155/2015/670691.
2. Shimokawa A., Kawasaki Y., Miyaoka E. Comparison of Splitting Methods on Survival Tree // Int. J. Biostat. 2015. Vol. 1, № 11. P. 175–188. DOI: 10.1515/ijb-2014-0029.
3. Comprehensive molecular portraits of human breast tumours [Electronic resource]. – Mode of access: https://www.nature.com/articles/nature11412. – Date of access: 24.12.2021.
4. Ishwaran H., Kogalur U. B., Blackstone E. H., Lauer M. S. Random survival forests // Ann. Appl. Stat. 2008. Vol. 2, № 12. P. 841–860. DOI: 10.1214/08-AOAS169.
5. Hornung R., Wright M. N. Block Forests: random forests for blocks of clinical and omics covariate data // BMC Bioinformatics. 2021. Vol. 22. DOI: 10.1214/08-AOAS169.