ПРЕДСКАЗАНИЕ ВЫЖИВАЕМОСТИ ПАЦИЕНТОВ С ОНКОЛОГИЧЕСКИМИ ЗАБОЛЕВАНИЯМИ МЕТОДОМ СЛУЧАЙНОГО ЛЕСА

В. Н. Яцков, М. К. Чепелева

Белорусский государственный университет, г. Минск
vlad18742@gmail.com, maryna.chepeleva@gmail.com
науч. рук. - П. В. Назаров, канд. физ.-мат. наук, доц.

Разработаны программные средства для предсказания выживаемости пациентов с онкологическими заболеваниями на основе метода случайного леса с расщеплением узлов по методу exponential log-likelihood loss и определения значимых признаков. Получена точность 91,4 % для глиобластомы. Разработанный алгоритм может быть использован для предсказания клинических рисков в персонализированной медицине.

Ключевые слова: случайный лес выживаемости; секвенирование; предсказание выживаемости; отбор признаков.

ВВЕДЕНИЕ

Предсказание рисков для пациентов является одной из важных задач биомедицины. Особенно актуальной такая задача становится с распространением подхода персонализированной медицины (personalized medicine), который предполагает индивидуальное рассмотрение многомерных данных о конкретном пациенте и включает предсказание выживаемости [1]. Предсказание выживаемости подразумевает оценку времени наступления критического события, основанную на функции вероятности наступления события. Точное прогнозирование клинического риска необходимо для принятия решений об оптимальных планах лечения. В разработке лекарств важную роль играет определение мишеней для терапии – компонент в клетках, изменив которые, можно повлиять на течение болезни. Одним из инструментов для определения таких мишеней может служить оценка значимости признаков прогнозной модели высокой точности. Метод случайного леса отличается относительно простой установкой связи между ковариатами и риском, а также показывает высокую точность прогнозирования [2]. Цель данной работы – разработка программных средств для предсказания выживаемости пациентов, больных раком, на основе метода случайного леса выживаемости (random survival forest) и определения значимых признаков.

МАТЕРИАЛЫ И МЕТОДЫ

Для тестирования алгоритмов были использованы данные секвенирования РНК (RNA-seq) глиобластомы человека из баз данных TCGA (portal.gdc.cancer.gov/) и IVY GAP (https://glioblastoma.alleninstitute.org/). Объединенные данные содержали 508 образцов, для 452 наступило критическое событие в определенный момент времени, 56 – цензурированы. Помимо экспрессии генов имелся набор клинических признаков: пол, возраст, подтип рака, набор данных (dataset), метилирование промотора гена MGMT, амплификация гена эпидермального фактора роста EGFR.

Случайный лес выживаемости. 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 [3].

В данной работе алгоритм случайного леса выживаемости был реализован на основе R-пакета *rpart* с программированием собственного алгоритма расщепления узлов.

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

$$\hat{\lambda}_h = \frac{\sum_{l_i \in L_h} \delta_i}{\sum_{l_i \in L_h} t_i}.$$

Само значение EL в узле h определяется как

$$R(h) = \sum_{l_i \in L_h} \delta_i - \sum_{l_i \in L_h} \delta_i \ln\left(\hat{\lambda}_h\right). \quad (1)$$

Каждому расщеплению ставится в соответствие «добротность» – величина, обратно пропорциональная сумме значений (1) дочерних узлов. Расщепление, которое максимизирует «добротность», считается наилучшим.

Качество предсказания. Для оценки ошибки предсказания выживаемости вычислялся *C*-индекс (concordance index), который оценивает вероятность того, что в i-ой паре объектов выполняется условие $S_{i1}(t_{i1}) < S_{i2}(t_{i2})$

при $t_{i1} < t_{i2}$ [3], где $S(t)$ – вероятность выживания объекта на момент времени t. В результате полностью достоверного предсказания значение шибки по *C*-индексу равно 0, в результате полностью недостоверного – 1.

Оценка значимости признаков. Для определения значимости признаков деревьев, размещенных в объектах *rpart*, требуется задать количество суррогатных разбиений, такие разбиения используются только для оценки значимости признаков. При этом «добротности» суррогатных разбиений должны быть меньше «добротности» основного разбиения. Для каждого суррогатного разбиения вычисляется adjusted agreement – коэффициент совпадения назначений объектов родительским узлам по данному разбиению по отношению к назначениям объектов родительским узлам по основному разбиению. Adjusted agreement изменяется от 0 до 1, где 1 – полное совпадение результатов данного суррогатного разбиения с лучшим.

Значимость признака в дереве определена как сумма всех «добротностей» основных разбиений и значимостей суррогатных разбиений, где использовался данный признак. При этом значимость суррогатного разбиения вычисляется как «добротность» лучшего разбиения на данном узле, умноженная на adjusted agreement. Для получения значимостей признаков леса, результаты по деревьям усредняются.

РЕЗУЛЬТАТЫ

Для предсказания выживаемости пациентов использовались реализованный алгоритм случайного леса выживаемости на основе R-пакета *rpart*, случайный лес выживаемости из пакета *randomForestSRC*, алгоритм регрессии Кокса из пакета *survival*. Сравнение точности предсказаний представлены в таблице 1, где N – количество деревьев.

Таблица 1

Общие результаты работы алгоритмов

Алгоритм	Ошибка предсказания по *C*-индексу, %
Лес на основе *rpart*, $N = 10$	9,70
Лес на основе *rpart*, $N = 50$	8,63
Лес на основе *rpart*, $N = 100$	8,87
Лес из *randomforestSRC*, $N = 10$	12,42
Лес из *randomforestSRC*, $N = 50$	10,99
Лес из *randomforestSRC*, $N = 100$	10,95
Модель Кокса	33,25

Реализованный алгоритм на основе *rpart* показывает точность в среднем на 2,4 % лучше, чем соответствующий случайный лес выживаемости из пакета *randomforestSRC*.

В таблице 2 представлены 12 наиболее значимых признаков.

Таблица 2

Полученные значимости признаков

Название признака	Значимость, 10^{-2}	Название признака	Значимость, 10^{-2}
Подтип рака	0,816	LARP4	0,114
Амплификация EGFR	0,414	NDUFA11	0,113
Возраст	0,325	ATP6V1F	0,113
Пол	0,245	TRPV1	0,111
Набор данных	0,241	FNDC3B	0,110
Метилирование MGMT	0,134	CCDC142	0,109

Наибольшую значимость имеют категориальные признаки, что коррелирует с исследованиями в данной области. От молекулярного подтипа (subtype) зависит траектория развития опухоли и применяемая терапия. EGFR запускает механизмы, приводящие к делению клеток. При отсутствии метилирования промотора MGMT длительная выживаемость очень редка [4]. Отсутствие видимой корреляции функциональных продуктов значимых генов с развитием опухоли свидетельствует о необходимости анализа совокупности значимых генов с целью выявления измененных биологических функций.

ЗАКЛЮЧЕНИЕ

Разработаны программные средства для предсказания выживаемости пациентов методом случайного леса с расщеплением узлов на основе EL. При сравнении с методом *randomForestSRC* и регрессией Кокса достигнута наибольшая точность (91,4 % при 50 деревьях). Определены наиболее значимые признаки, среди которых подтип рака, амплификация EGFR, метилирование MGMT. Разработанный алгоритм 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. 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.
4. Smrdel U., Popovic M., Zwitter M., et al. Long-term survival in glioblastoma: methyl guanine methyl transferase promoter methylation as independent favourable prognostic factor // Radiol Oncol. 2016. Vol. 50, №4. P. 394–401. DOI:10.1515/raon-2015-0041.