ИССЛЕДОВАНИЕ ВЛИЯНИЯ МЕТОДОВ ПОНИЖЕНИЯ РАЗМЕРНОСТИ ДАННЫХ НА ТОЧНОСТЬ РАБОТЫ АЛГОРИТМА ПРЕДСКАЗАНИЯ ВЫЖИВАЕМОСТИ

В. Н. Яцков

Белорусский государственный университет, г. Минск;
vlad18742@gmail.com;
науч. рук.: М. К. Чепелева, ассист.

Использованы методы понижения размерности данных для сокращения вычислительной сложности и повышения эффективности работы алгоритма для предсказания выживаемости пациентов с онкологическими заболеваниями методом блочного леса с расщеплением узлов по методу *exponential log-likelihood loss*. Проведено сравнение качества предсказания при снижении размерности методами главных компонент и независимых компонент. Получена лучшая точность 84,55 % по Бриеру для блочного леса при использовании метода независимых компонент. Данная работа может быть использована для улучшения программных средств предсказания клинических рисков в персонализированной медицине.

Ключевые слова: блочный лес; секвенирование; предсказание выживаемости; метод главных компонент; метод независимых компонент.

ВВЕДЕНИЕ

Современный уровень развития биомедицинских технологий позволяет при лечении заболеваний использовать подход персонализированной медицины (personalized medicine) – индивидуальное рассмотрение данных конкретного пациента для принятия решений [1]. Такой подход является актуальным при онкологических заболеваниях вследствие высокой гетерогенности опухолей, и одной из его составляющих является предсказание рисков для пациентов.

Предсказание выживаемости является оценкой времени наступления критического события, основанной на функции вероятности наступления данного события. Прогнозирование функции выживаемости пациента и установление влияния признаков (в число которых входят различные виды терапии) позволяют принять решение об оптимальном плане лечения.

Данная работа подразумевает анализ многомерных данных. Однако предсказание выживаемости сразу на исходных данных (экспресссии генов) не рационально по двум причинам. Во-первых, получается слишком высокая вычислительная сложность. Во-вторых, не все признаки имеют предсказательную способность.

Цель данной работы – разработка алгоритма для предсказания выживаемости пациентов, больных раком молочной железы, на основе метода блочного леса, и повышение эффективности его работы за счет использования методов снижения размерности данных.

МАТЕРИАЛЫ И МЕТОДЫ

Для тестирования алгоритмов были использованы данные секвенирования РНК для пациентов с раком груди [2]. 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 20119 признаков, описывающих экспрессию генов, имелись клинические признаки: пол, подтип рака, тип образца ткани, группа по наличию раковой опухоли.

Сокращение размерности данных

В данной работе были использованы два метода для уменьшения числа признаков: метод главных компонент и метод независимых компонент.

Метод главных компонент (МГК) представляет собой вращение системы координат исходных данных с целью максимизировать дисперсию новых признаков [3].

Метод независимых компонент (МНК) – метод разложения линейных сигналов на аддитивные независимые компоненты. При этом мерой независимости в [4] считается наибольшее отклонение от гауссовости. Аппроксимация данного выражения:

$$J \approx \sum_{j=1}^{k} \left[E(G(y_i)) - E(G(y_j^{gauss})) \right]^2 \quad (1)$$

где $y_i = w_i x$, x – вектор исходных данных размера k, w_i – i-я строка матрицы преобразования исходных данных в независимые компоненты, y_j^{gauss} – вектор случайных гауссовых значений с параметрами, как у y_j, E – матожидание, G – одна из заданных функций:

$$G(u) = \frac{1}{a}\log(\cosh(au)), a \in [1,2] \quad (2)$$

$$G(u) = -\exp\left(-\frac{u^2}{2}\right) \quad (3)$$

Предсказание выживаемости

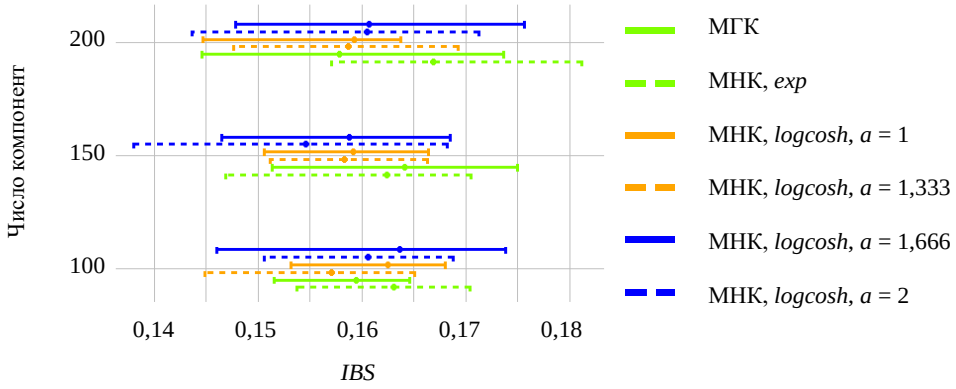
Основываясь на ранее проведенном исследовании по сравнению точности алгоритмов предсказания выживаемости [5], был выбран блочный лес с расщеплением узлов по методу *exponential log-likelihood loss*, также описанным в [5], поскольку он показал лучший результат среди сравниваемых алгоритмов.

Качество предсказания

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 была предсказана алгоритмом блочного леса с *EL* расщеплением узлов. Материал для предсказания был получен следующим образом: исходные данные были обработаны названными выше методами снижения размерности при 100, 150 и 200 компонентах, причем для МНК были взяты функции *logcosh* (2) и *exp* (3) для аппроксимации (1). Для (2) были взяты значения параметра, описанные на рисунке . На рисунке показаны распределения ошибок предсказания выживаемости методом блочного леса, но с различными методами и параметрами уменьшения размерности данных.



По полученным результатам можно сделать вывод, что использование МНК с функцией *logcosh* для аппроксимации (1) дает лучший результат, чем МГК для уменьшения размерности. МНК с функцией *exp* для аппроксимации (1) не дает определенного результата.

ЗАКЛЮЧЕНИЕ

Было проведено сравнение методов понижения размерности данных: метод главных компонент и метод независимых компонент. Алгоритм блочного леса показал лучшую среднюю оценку 84,55 % по оценке Бриера на данных, полученных от МНК. Данный результат может быть использован для улучшения работы программных средств 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. Comprehensive molecular portraits of human breast tumours [Electronic resource] – Mode of access: https://www.nature.com/articles/nature11412. – Date of access: 24.12.2021.
3. *Яцков, Н. Н.* Интеллектуальный анализ данных / *Н. Н. Яцков.* – Минск : БГУ, 2014. – 151 с.
4. *Shchurenkova, E.* Dimension reduction using Independent Component Analysis with an application in business psychology : master of science / *E. Shchurenkova*. – Vancouver, 2017. – 72 p
5. Разработка алгоритма предсказания выживаемости пациентов с онкологическими заболеваниями : материалы III Междунар. науч.-практ. конф., Минск, 10–21 апр. 2022 г. / Белорус. гос. ун-т ; редкол.: В. В. Скакун (отв. ред) [и др.]. – Минск : БГУ, 2022. – 317 с.