Секция «Биоинформатика»

IndieForest: машинное обучение на индикаторах и рангах для ранней диагностики онкогенных заболеваний

Научный руководитель – Фаворов Александр Владимирович

Гурылева Мария Вячеславовна
Студент (специалист)
Московский государственный университет имени М.В.Ломоносова, Факультет биоинженерии и биоинформатики, Москва, Россия
E-mail: guryleva.mv@gmail.com

Онкологические заболевания характеризуются высокой скоростью прогрессирования. Поэтому особенно важна их ранняя диагностика. Методы, используемые для диагностики сейчас, сильно зависят от человеческого фактора, а также ограничены разрешающей способностью приборов. Анализ экспрессий генов может помочь детекции рака на ранних стадиях. Для такой задачи наиболее подходящим подходом видится использование алгоритмов машинного обучения.[1]

Один из таких алгоритмов - случайный лес (RF) - хорошо зарекомендовал себя в биоинформатических задачах.[2] Лес — это голосование большого числа решающих деревьев, каждое из которых строится на случайной подвыборке образцов и переменных. Решающее дерево состоит из набора элементарных решений, которые сравнивают значения переменных с порогами. Пороги подбираются при обучении и становятся частью классификатора. Из-за этого классифицируемые данные экспрессии надо нормализовать вместе с обучающей выборкой, что ограничивает применимость RF.

В то же время существует семейство непараметрических методов, основанных на попарных сравнениях экспрессий генов внутри одного образца, что делает их независящими от монотонной нормализации.[3]

Цель данного проекта — соединить два подхода и использовать для предсказания различных типов рака RF, обученный на результатах попарных сравнений экспрессий генов образца.

Данная идея была реализована на языке программирования R с использованием пакета randomFosest и набора пакетов для обработки данных tidyverse. Тестирование моделей проводилось на данных из баз данных TCGA (https://portal.gdc.cancer.gov/) и GEO (https://www.ncbi.nlm.nih.gov/geo/). Для избежания переобучения отбирались наиболее вариабельные дифференциально экспрессирующиеся гены.

Далее рассматривались три метода: классический - основанный на стандартных показателях экспрессии генов, ранговый - показатели экспрессии переводились в ранги, метод индикаторов - в качестве признака использовалась величина I[A, B], равная 1 если $A > B$ и -1 иначе. Сравнение результатов для бинарной классификации проводилось относительно алгоритма K-Top-Scoring-Pair (KTSP), реализованного в пакете switchBox.[3]

На бинарной классификации было показано, что уже на 100 генах и индикаторный, и ранговый методы превосходят алгоритм KTSP по таким показателям, как accuracy и ROC-auc. Было показано, что данные методы применимы и для множественной классификации различных типов рака, в отличие от стандартного KTSP.

Источники и литература

1) Ming, C., Viassolo, V., Probst-Hensch, N. et al. Machine learning techniques for personalized breast cancer risk prediction: comparison with the BCRAT and BOADICEA models. Breast Cancer Res 21, 75 (2019). https://doi.org/10.1186/s13058-019-1158-4

2) Breiman, L., 2001. Random Forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324

3) Afsari, B., Fertig, E.J., Geman, D., Marchionni, L., 2015. switchBox: an R package for k–Top Scoring Pairs classifier development. Bioinformatics 31, 273–274. https://doi.org/10.1093/bioinformatics/btu622