УДК 004.89

МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ В ЗАДАЧЕ ГЕНЕРАЦИИ РЕАЛИСТИЧНЫХ СИНТЕТИЧЕСКИХ МЕТАГЕНОМНЫХ НАБОРОВ ДАННЫХ

Иванов А.Б. (Университет ИТМО)
Научный руководитель – канд. техн. наук Ульянцев В.И. (Университет ИТМО)

Введение.

Микробные сообщества населяют различные ниши окружающего мира, такие как водоемы, почва и организм человека. В человеке бактерии участвуют в переработке и усвоении питательных веществ, и регуляции иммунного ответа. Изучением всех микробов в совокупности из одного образца занимается метагеномика. Развитие современных методов полногеномного метагеномного секвенирования позволило извлекать всю информацию из метагеномного образца и изучать все микробное разнообразие и их взаимодействия, в то время как культивирование отдельных бактерий в лаборатории является нерешенной задачей для многих видов.

Изучение микробиоты кишечника является актуальной задачей, поскольку она играет важную роль в организме человека. Последние исследования показывают, что микробиом кишечника влияет на успешность иммунотерапии раковых заболеваний [1], развитие воспалительных заболеваний кишечника [2] и другие заболевания.

Для обработки данных разрабатываются различные алгоритмы, которые нуждаются в тестировании и валидации. При этом существует дефицит модельных микробных сообществ с заданными свойствами. Реальные метагеномные данные не подходят для этих целей, так как в них достоверно неизвестен ни таксономический состав, ни абсолютные значения представленности бактерий. Поэтому стоит задача создания синтетических метагеномных наборов данных с фиксированными свойствами, близкими к реальным.

Основная часть.

Для решения задачи генерации метагеномных образцов был разработан метод samovaR (https://github.com/dsmutin/samovar). Он позволяет генерировать вектора представленности метагеномных видов, из которых затем с помощью программы InSilicoSeq [3] могут быть получены файлы с прочтениями в формате FASTQ, как от реальных секвенаторов платформы Illumina. Для генерации образцов необходима информация о реальных метагеномах, которая извлекается из базы данных GMRepo [4].

В данной работе подход samovaR оценивается и валидируется про помощи состязательных нейронных сетей [5]. Во-первых, производится кластеризация, чтобы проверить разницу между векторами реальных примеров из базы данных и сгенерированными векторами.

Во-вторых, проводится обучение нейронной сети для поиска отличий реальных векторов от случайно сгенерированных векторов с белым шумом. Затем она тестируется на векторах, полученных в результате работы алгоритма samovaR. Поскольку целью является генерация реалистичных метагеномов, ожидаемым результатом будет похожесть сгенерированных метагеномов на настоящие.

В-третьих, разрабатывается архитектура генеративной состязательной нейронной сети для создания метагеномных векторов представленности видов. Она позволит создавать образцы, отталкиваясь от случайного нормального распределения. Проводится оценка наилучшего метода генерации метагеномов со свойствами, наиболее близкими к реальным данным.

Реализация моделей машинного обучения производится на языке программирования python3 с использованием библиотеки PyTorch.

Выводы.

В данной работе применены алгоритмы глубокого машинного обучения для генерации и валидации синтетических наборов метагеномных данных со свойствами, имитирующими реальные образцы.

Список использованных источников:

1. Frankel A. E. et al. Metagenomic shotgun sequencing and unbiased metabolomic profiling identify specific human gut microbiota and metabolites associated with immune checkpoint therapy efficacy in melanoma patients //Neoplasia. – 2017. – Т. 19. – №. 10. – С. 848-855.
2. Franzosa E. A. et al. Gut microbiome structure and metabolic activity in inflammatory bowel disease //Nature microbiology. – 2019. – Т. 4. – №. 2. – С. 293-305.
3. Gourlé H. et al. Simulating Illumina metagenomic data with InSilicoSeq //Bioinformatics. – 2019. – Т. 35. – №. 3. – С. 521-522.
4. Wu S. et al. GMrepo: a database of curated and consistently annotated human gut metagenomes //Nucleic acids research. – 2020. – Т. 48. – №. D1. – С. D545-D553.
5. Goodfellow I. et al. Generative adversarial nets //Advances in neural information processing systems. – 2014. – Т. 27.

Иванов А.Б. (автор)

Ульянцев В.И. (научный руководитель)