УДК 004.891.2

МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ В ЗАДАЧАХ СРАВНИТЕЛЬНОЙ МЕТАГЕНОМИКИ С ПРИМЕНЕНИЕМ ОМИКСНЫХ ДАННЫХ

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

В данной работе описаны методы для извлечения признаков из метагеномных данных. Проводятся масштабные вычислительные эксперименты по классификации метагеномных образцов с применением методов машинного обучения. Предложен алгоритм по использованию метатранскриптомных данных для отбора наиболее значимых признаков.

Введение.

Микробные сообщества населяют различные экосистемы окружающего мира, в том числе организм человека, где они играют важную роль в усвоении питательных веществ и регуляции иммунного ответа. Метагеномика занимается анализом таких сообществ, а именно определением микроорганизмов, которые есть в сообществе, и их функциональной роли, и установлением взаимосвязей, которые существуют между бактериями, а также между сообществом и окружающей средой. Полногеномное метагеномное секвенирование позволяет исследовать сообщество в целом и получить информацию о населяющих его микроорганизмах, которые могут быть некультивируемы по отдельности. Развитие метагеномики и методов секвенирования нового поколения привело к накоплению большого массива метагеномных данных, анализ которых невозможен без современных вычислительных методов. В то же время существующие методы не вполне справляются с этой задачей в связи с объемом, комплексностью и зашумленностью данных.

Классическим методом сравнительного анализа метагеномов является извлечение из образцов информации о том, какие известные микроорганизмы содержатся в них (таксономическая аннотация) и какие метаболические функции они выполняют (функциональная аннотация). Недостатком данных подходов является возможность анализировать только ту часть образца, информация о которой находится в базах данных, что может приводить к потере существенного объема данных и пропуску ключевых биологических факторов. Другая группа методов позволяет анализировать образцы на основе коротких последовательностей k-меров. Это позволяет исследовать весь объем данных, однако получаемые результаты плохо поддаются биологической интерпретации.

Помимо анализа метагеномных данных, популярность набирают другие методы анализа исследуемых образцов с целью получения разносторонней информации. Так, транскриптомика позволяет определить, какие клеточные процессы были активны и каки гены транскрибировались с последовательности ДНК. Протеомика позволяет обнаруживать и выделять белки, которые синтезировались в клетках. Их анализ совместно с метагеномными данными может позволить получить более полную информацию об исследуемых микробных сообществах.

Основная часть.

В данной работе исследуется применимость методов извлечения метагеномных признаков и машинного обучения для решения задач сравнительной метагеномики. Ранее нами был предложен алгоритм для извлечения признаков из метагеномных наборов данных на основе отбора уникальных k-меров и построения графов де Брейна. Такой подход позволяет не только проанализировать полную информацию, доступную об образце в результате секвенирования, но и повышает вероятность биологической интерпретации полученных результатов за счет работы с признаками в виде относительно длинных геномных последовательностей. Данный алгоритм был применен для анализа метагеномных образцов секвенирования микробиоты кишечника людей с заболеваниями желудочно-кишечного тракта, а именно с болезнью Крона или язвенным колитом.

В настоящий момент данные заболевания слабо поддаются ранней диагностики и требуют инвазивных процедур для подтверждения заболевания. Использование данных о микробиоте кишечника является перспективным направлением в диагностике воспалительных заболеваний кишечника (ВЗК). Нами предлагается следующая процедура анализа данных метагеномного секвенирования для предсказания ВЗК у пациентов. В качестве входных данных алгоритм принимает данные метагеномного секвенирования образцов, предварительно отфильтрованные от возможных загрязнений и ошибок, разделенные на категории по заболеваниям. Для каждой категории производится выделение уникальных k-меров, которые затем объединяются в протяженные участки ДНК (контиги). Извлечение контигов может производиться как с помощью разрабатываемого алгоритма MetaFast, так и с помощью сторонних метагеномных сборщиков (например, metaSpades). Далее каждый контиг выступает отдельным признаком, числовое значение которого вычисляется как доля его покрытия k-мерами из образца. Полученная матрица признаков используется в алгоритмах машинного обучения для классификации метагеномных образцов.
В данной работе были применены различные модели машинного обучения, такие как случайные лес деревьев решений, полносвязные нейронные сети и алгоритмы бустинга. Для вычислительных экспериментов использовались четыре набора данных, на одном из которых проводилось обучение моделей, а три других использовались для оценки их качества. Результаты точности классификации с использованием извлеченных признаков превышают значения, достигаемые при использовании информации о таксономической или функциональной аннотации образцов для всех наборов данных. При этом выбор алгоритма машинного обучения с наилучшим значением качества и подбор его параметров зависит от исследуемых данных.
Предложенный алгоритм извлечения признаков выдает большое число признаков (десятки тысяч), что затрудняет обучение классификационных моделей. Для решения этой проблемы предлагаются два способа.
Первые способ состоит в использовании алгоритмов, оценивающих вклад каждого признака в обученную модель и позволяющих отобрать подмножество признаков, действительно важных для правильной классификации образцов. Эксперименты показали, что применение таких методов позволяет значительно сократить число признаков и ускорить обучение моделей с незначительной потерей в качестве результатов классификации.
Второй способ направлен на использование метатранскриптомных данных для отбора наиболее значимых из извлеченных признаков. Был реализован алгоритм, который позволяет отбирать только те контиги, которые хорошо покрываются метатранскриптомными прочтениями. С биологической точки зрения это означает, что отобранные контиги соответствуют участкам ДНК, с которых экспрессируется информация в исследуемом образце. Предварительные эксперименты показывают, что такая фильтрация позволяет повысить точность классификации. В дальнейшем планируется развитие данных подходов и внедрение других видов омиксных данных.

Выводы. В данной работе были проведены вычислительные эксперименты по использованию методов извлечения метагеномных признаков и алгоритмов машинного обучения для классификации воспалительных заболеваний кишечника. Был предложен метод использования метатранскриптомных данных для повышения точности классификации. Предлагаемые подходы могут служить основой для разработки системы поддержки принятия решений при диагностировании заболеваний на основе метагеномных данных.

Иванов А.Б. (автор) _____

Ульянцев В.И. (научный руководитель) _____