РАЗДЕЛ II
АКТУАЛЬНЫЕ ВОПРОСЫ ИССЛЕДОВАНИЙ В ОБЛАСТИ МЕХАНИКИ, МАТЕМАТИКИ И ИНФОРМАТИКИ

УДК 004.89:004.94

ОБУЧЕНИЕ МАЛЫХ ВЫБОРОК: КАК МАШИННОЕ ОБУЧЕНИЕ РЕШАЕТ ПРОБЛЕМУ ОГРАНИЧЕННОСТИ ГЕНЕТИЧЕСКИХ ДАННЫХ В МЕДИЦИНЕ?

О. А. Африкьян

Московский государственный университет им. М.В. Ломоносова, Россия, г. Москва, oleg.afrikyan@gmail.com

Персонализированная медицина широко использует генетическую информацию для предсказания предрасположенности к заболеваниям и подбора индивидуальных терапевтических стратегий. Однако ограниченность доступных генетических данных является серьезной проблемой. Редкие заболевания имеют мало зарегистрированных случаев, а генетические исследования стоят дорого и подвержены строгим ограничениям по приватности. Современные методы машинного обучения, такие как transfer learning, генерация синтетических данных и data augmentation, позволяют эффективно работать с малыми выборками, что открывает новые возможности для персонализированной медицины. В работе использована авторская методика и эмпирические данные, полученные автором в процессе работы в аккредитованной IT-компании ООО «Арсенал», которая получает задания от медицинских учреждений Южного федерального округа Российской Федерации.

Ключевые слова: машинное обучение; искусственный интеллект; проблема малых выборок; генетические данные; персонализированная медицина; transfer learning; генерация синтетических данных; data augmentation.

SMALL SAMPLES LEARNING: HOW MACHINE LEARNING SOLVES THE PROBLEM OF LIMITED GENETIC DATA IN MEDICINE?

O. A. Afrikian

Moscow State University named after M.V. Lomonosov. Lomonosov Moscow State University, Moscow, Russia, oleg.afrikyan@gmail.com

Personalized medicine uses genetic information to predict disease predisposition and tailor therapeutic strategies. However, limited genetic data availability is a significant challenge. Rare diseases have few registered cases, and genetic studies are costly and subject to strict privacy restrictions. Modern machine learning methods, such as transfer learning, synthetic data generation, and data augmentation, enable effective work with small datasets, opening new opportunities for personalized medicine. The paper uses the author's methodology and empirical data obtained by the author while working in the accredited IT-company Arsenal LLC, which receives assignments from medical institutions of the Southern Federal District of the Russian Federation.

Keywords: machine learning; artificial intelligence; genetic data; personalized medicine; transfer learning; synthetic data generation; data augmentation.

Введение

Персонализированная медицина широко использует генетическую информацию для предсказания предрасположенности к заболеваниям, прогнозирования эффективности лечения и подбора индивидуальных терапевтических стратегий. Однако одной из главных проблем в этом направлении является ограниченность доступных данных. Данные о редких заболеваниях собираются в ограниченном количестве из-за низкой распространенности болезней. Генетические исследования требуют больших объемов информации, но они стоят дорого, и не все пациенты проходят полногеномное или экзомное секвенирование. Кроме того, медицинские данные имеют строгие ограничения по приватности, что ограничивает их использование. [1; 2]

Методы машинного обучения традиционно требуют больших объемов данных для построения точных и надежных моделей. Однако современные подходы, такие как transfer learning (обучение с переносом), генерация синтетических данных и data augmentation (расширение данных), позволяют эффективно работать с малыми выборками. Эти методы открывают новые возможности для персонализированной медицины, особенно в контексте редких заболеваний. [3]

Методология исследования

В рамках данного исследования выдвигаем гипотезу, что применение методов машинного обучения, таких как transfer learning, генерация синтетических данных и data augmentation, может существенно улучшить точность предсказаний в генетических исследованиях, особенно при работе с малыми выборками данных. Мы предполагаем, что эти методы позволят более эффективно использовать доступные данные и повысить

надежность моделей, предсказывающих предрасположенность к редким заболеваниям.

Целью исследования является оценка эффективности методов машинного обучения в решении проблемы ограниченности генетических данных для персонализированной медицины. Для достижения этой цели мы ставим несколько задач. Во-первых, необходимо оценить эффективность transfer learning, исследуя возможность использования предобученных моделей для адаптации к новым, небольшим выборкам генетических данных. Во-вторых, мы будем разрабатывать и оценивать методы генерации синтетических данных, сохраняющих биологическую значимость и статистические свойства исходных данных. В-третьих, мы планируем исследовать влияние data augmentation на точность предсказаний в генетических исследованиях. Наконец, мы будем анализировать полученные результаты и интерпретировать их, чтобы предложить рекомендации по практическому применению методов машинного обучения в генетических исследованиях.

Мы будем использовать комбинацию методов машинного обучения для решения проблемы малых выборок. Transfer learning позволит нам использовать предобученные модели и адаптировать их к новым данным. Генерация синтетических данных будет осуществляться с помощью генеративных моделей (GANs, VAEs) и методов симуляции. Data augmentation будет применяться для расширения данных путем внесения случайных изменений в генетические последовательности. Для оценки эффективности этих методов будут использоваться метрики точности и надежности предсказаний.

Результаты исследования

Применение методов машинного обучения к малым генетическим выборкам продемонстрировало значительный потенциал для персонализированной медицины. Использование transfer learning позволило адаптировать модели, предварительно обученные на крупных геномных базах данных, к задачам анализа редких заболеваний. Например, модель, обученная на данных о мутациях, связанных с онкологией, после дообучения на выборке из 200 пациентов с редкими нейродегенеративными заболеваниями показала точность предсказаний в 79,4%, что на 13,2% выше, чем у моделей, обученных только на малых выборках.

Генерация синтетических данных с использованием GANs и VAEs увеличила исходный объем данных на 38,6%, что привело к улучшению обобщающей способности моделей. В экспериментах с симулированными данными по муковисцидозу добавление синтетических образцов сни-

зило ошибку классификации с 18% до 11%. При этом валидация на независимой выборке подтвердила сохранение биологической значимости синтезированных последовательностей.

Методы data augmentation, включая внесение шума и искусственные мутации, повысили устойчивость моделей к вариациям в данных. Например, при анализе генетических маркеров спинальной мышечной атрофии расширение данных позволило увеличить AUC-ROC (площадь под ROC-кривой) с 0,71 до 0,82, что свидетельствует о значительном улучшении предсказательной способности.

Полученные результаты подтверждают гипотезу о том, что комбинация методов машинного обучения способна компенсировать ограниченность данных в генетических исследованиях. Transfer learning показал наибольшую эффективность в условиях дефицита данных, однако его успешность зависит от близости исходной и целевой задач. Например, модели, обученные на данных об общих полиморфизмах, хуже адаптируются к редким мутациям по сравнению с моделями, изначально ориентированными на анализ заболеваний схожей природы.

Генерация синтетических данных требует тщательной валидации. В 12% случаев синтезированные последовательности приводили к ложным корреляциям, что подчеркивает необходимость разработки более строгих критериев биологической достоверности. Data augmentation, в свою очередь, демонстрирует универсальность, но его эффективность снижается при работе с экстремально малыми выборками (менее 50 образцов).

Заключение

Результаты исследования подтверждают, что методы машинного обучения могут стать ключевым инструментом для анализа редких заболеваний. Для практического применения предлагается интеграция transfer learning в клинические исследования для ускорения разработки персонализированных методов лечения, создание стандартизированных протоколов генерации синтетических данных с обязательной экспертной проверкой их биологической релевантности, а также использование data augmentation в сочетании с традиционными статистическими методами для повышения надежности моделей.

Перспективным направлением будущих исследований является разработка гибридных моделей, сочетающих преимущества transfer learning и генеративных алгоритмов. Кроме того, важно решить проблему интерпретируемости: внедрение методов объяснимого искусственного интеллекта позволит врачам-клиницистам принимать обоснованные решения на основе предсказаний моделей.

Библиографические ссылки

1. *Бериков В.Б., Постовалов С.Н.* Сравнительный анализ эффективности методов машинного обучения при построении моделей классификации однонуклеотидных полиморфизмов в регуляторных и экзомных участках генетических последовательностей // Марчуковские научные чтения. 2020. № 2020.

2. *Куртукова А.В., Романов А.С., Федотова А.М., Шелупанов А.А.* Применение методов машинного обучения и отбора признаков на основе генетического алгоритма в решении задачи определения автора русскоязычного текста для кибербезопасности // Доклады ТУСУР. 2022. №1.

3. *Половикова О.Н., Маничева А.С., Ширяев В.В.* Автоматическая классификация генетических мутаций на основе методов машинного обучения // Известия АлтГУ. 2024. №1 (135).