МЕТОДЫ И АЛГОРИТМЫ АНАЛИЗА И ОБРАБОТКИ ОДНОКЛЕТОЧНЫХ ОМИКСНЫХ ДАННЫХ НА ОСНОВЕ ОБУЧЕНИЯ НЕЙРОННЫХ СЕТЕЙ

Мусакаев Руслан
Ташкентский Государственный технический университет
Аспирант (PHD)
Email: marasres08121999@mail.ru

Аннотация

В докладе рассматриваются современные методы анализа и обработки одноклеточных омиксных данных с применением алгоритмов глубокого обучения. Особое внимание уделяется архитектурам нейронных сетей, включая автоэнкодеры, вариационные модели, графовые нейронные сети и трансформеры, используемые для извлечения скрытых биологических закономерностей. Представлены алгоритмы снижения размерности, кластеризации, интеграции многомодальных данных, реконструкции клеточных траекторий и импутации пропущенных значений. Обсуждаются преимущества, ограничения и перспективы развития нейросетевых подходов в биоинформатике.

Annotatsiya

Ushbu ma’ruzada chuqur o‘rganish algoritmlaridan foydalangan holda bir hujayrali omiks ma’lumotlarini tahlil qilish va qayta ishlashning zamonaviy usullari ko‘rib chiqiladi. Asosiy e’tibor yashirin biologik qonuniyatlarni aniqlash uchun qo‘llaniladigan neyron tarmoqlar arxitekturalariga, jumladan avtoenkoderlar, variatsion modellar, graf neyron tarmoqlari va transformerlarga qaratilgan. O‘lchamni kamaytirish, klasterlash, ko‘p modalli ma’lumotlarni integratsiya qilish, hujayra trayektoriyalarini rekonstruksiya qilish hamda yetishmayotgan qiymatlarni tiklash algoritmlari taqdim etilgan. Shuningdek, bioinformatikada neyron tarmoqlarga asoslangan yondashuvlarning afzalliklari, cheklovlari va rivojlanish istiqbollari muhokama qilinadi.

Abstract

This report examines modern methods for the analysis and processing of single-cell omics data using deep learning algorithms. Particular attention is given to neural network architectures, including autoencoders, variational models, graph neural networks, and transformers, which are used to extract hidden biological patterns. Algorithms for dimensionality reduction, clustering, multimodal data integration, reconstruction of cellular trajectories, and imputation of missing values are presented. The advantages, limitations, and future prospects of neural network-based approaches in bioinformatics are also discussed.

1. Введение

Современные достижения в области высокопроизводительного секвенирования позволили перейти от анализа популяций клеток к исследованию отдельных клеток.

Одноклеточные омиксные технологии, такие как scRNA-seq, scATAC-seq и пространственная транскриптомика, формируют новый уровень понимания клеточной гетерогенности.

Однако такие данные характеризуются:

- высокой размерностью; разреженностью;
- значительным уровнем шума; нелинейными зависимостями.

Это делает традиционные методы анализа недостаточно эффективными и обуславливает необходимость применения методов глубокого обучения.

2. Постановка задачи

Пусть задан набор данных:

X={x1,x2,...,xn},xi∈RdX = \{x_1, x_2, ..., x_n\}, \quad x_i \in \mathbb{R}^dX={x1,x2,...,xn},xi∈Rd

где:

- nnn — число клеток, ddd — число признаков (генов, пиков, белков).

Требуется решить следующие задачи:

1. Снижение размерности: f:Rd→Rkf: \mathbb{R}^d \rightarrow \mathbb{R}^kf:Rd→Rk, k≪dk \ll dk≪d;
2. Кластеризация клеток;
3. Интеграция различных модальностей;
4. Реконструкция динамических процессов;
5. Импутация пропущенных значений.

3. Архитектуры нейронных сетей

3.1. Автоэнкодеры

Автоэнкодер представляет собой композицию функций:

z=fθ(x),x^=gϕ(z)z = f_\theta(x), \quad \hat{x} = g_\phi(z)z=fθ(x),x^=gϕ(z)

Функция потерь:

L=∥x−x^∥2\mathcal{L} = \|x - \hat{x}\|^2L=∥x−x^∥2

Варианты:

- Деноизирующие автоэнкодеры (DAE);
- Вариационные автоэнкодеры (VAE):

L=Eq(z|x)[log⁡p(x|z)]−KL(q(z|x)∥p(z))\mathcal{L} = \mathbb{E}_{q(z|x)}[\log p(x|z)] - KL(q(z|x)\|p(z))L=Eq(z|x)[logp(x|z)]−KL(q(z|x)∥p(z))

Применение: устранение шума; латентное представление; генерация данных.

3.2. Генеративные состязательные сети (GAN)

GAN состоят из генератора GGG и дискриминатора DDD:

min⁡Gmax⁡DE[log⁡D(x)]+E[log⁡(1−D(G(z)))]\min_G \max_D \mathbb{E}[\log D(x)] + \mathbb{E}[\log(1 - D(G(z)))]GminDmaxE[logD(x)]+E[log(1−D(G(z)))]

Используются для: генерации синтетических клеток; аугментации данных; борьбы с batch-эффектами.

3.3. Графовые нейронные сети (GNN)

Данные представляются графом:

G=(V,E)G = (V, E)G=(V,E)

Обновление признаков:

hi(k+1)=σ(∑j∈N(i)Whj(k))h_i^{(k+1)} = \sigma \left(\sum_{j \in \mathcal{N}(i)} W h_j^{(k)} \right)hi(k+1)=σj∈N(i)∑Whj(k)

Применение: кластеризация; анализ взаимодействий клеток; построение клеточных графов.

3.4. Трансформеры и механизмы внимания

Функция внимания:

Attention(Q,K,V)=softmax(QKTd)VAttention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right)VAttention(Q,K,V)=softmax(dQKT)V

Применяются для: моделирования глобальных зависимостей; анализа траекторий; интеграции данных.

4. Алгоритмы обработки данных

4.1. Алгоритм снижения размерности (на основе VAE)

Вход: матрица экспрессии XXX
Выход: латентное представление ZZZ

Шаги: Инициализация параметров сети; Обучение энкодера и декодера; Вычисление функции потерь (ELBO); Оптимизация методом Adam; Получение латентных переменных.

4.2. Алгоритм кластеризации

1. Получение латентного пространства;
2. Применение k-means или DBSCAN;
3. Оценка качества (Silhouette score);
4. Интерпретация кластеров.

4.3. Интеграция многомодальных данных

Подход:

- совместное обучение (joint embedding);
- выравнивание распределений;
- использование shared latent space.

4.4. Реконструкция траекторий

Алгоритм:

1. Построение графа ближайших соседей;
2. Вычисление псевдовремени;
3. Обучение модели последовательностей;
4. Восстановление переходов состояний.

4.5. Импутация данных

Методы: автоэнкодеры; VAE; diffusion-модели.

5. Экспериментальные аспекты

Метрики:

- ARI (Adjusted Rand Index); NMI (Normalized Mutual Information);
- Silhouette score; Reconstruction error.

Проблемы:

- переобучение; чувствительность к гиперпараметрам;
- вычислительная сложность.

6. Преимущества нейросетевых методов

- Моделирование нелинейных зависимостей;
- Высокая точность; Универсальность;
- Работа с большими данными; Интеграция модальностей.

7. Ограничения

- Слабая интерпретируемость; Требования к ресурсам;
- Необходимость больших выборок; Сложность настройки.

8. Перспективы развития

- Self-supervised learning; Foundation models в биологии;
- Интеграция пространственных данных; Explainable AI;
- Гибридные модели.

9. Заключение

Методы глубокого обучения открывают новые возможности для анализа одноклеточных омиксных данных. Они обеспечивают эффективное извлечение скрытых закономерностей, повышают точность кластеризации и позволяют моделировать динамику клеточных процессов. Несмотря на ограничения, данные методы являются перспективным направлением развития современной биоинформатики.