ГЕНЕРАТИВНОЕ ИМИТАЦИОННОЕ МОДЕЛИРОВАНИЕ СЛОЖНЫХ БИОФИЗИЧЕСКИХ СИСТЕМ

Н.Н. Яцков, В.В. Апанасович, В.Н. Яцков

Белорусский государственный университет, Минск, Беларусь
E-mail: yatskou@bsu.by

Разработан метод генеративного имитационного моделирования для исследования биофизических систем, информация о поведении которых представлена в виде больших наборов экспериментальных данных. Предложенный метод является инструментом искусственного интеллекта и основан на генерации адекватных сценариев поведения систем, используя имитацию случайных событий, вероятностные распределения которых находятся из имеющихся экспериментальных данных. Применение метода позволяет повысить эффективность анализа сложных биофизических систем.

Ключевые слова: искусственный интеллект; генеративное моделирование; имитационная модель; машинное обучение; биофизическая система.

ВВЕДНИЕ

Использование методов имитационного моделирования в ходе анализа больших данных открывает новые возможности в области высокотехнологичных междисциплинарных исследований [1]. Имитационное моделирование позволяет осуществлять анализ систем практически любой сложности и проводить вычислительные эксперименты, труднодостижимые в натурных исследованиях [1].

В современных научных исследованиях наблюдается тенденция применения генеративного моделирования, целью которого является создание реалистичных и разнообразных образцов данных по заданным распределениям [2]. Генеративная модель описывает, как генерируется набор данных, с точки зрения вероятностной модели. Используя эту модель, можно генерировать новые данные [3].

Генеративное моделирование может использоваться в биоинформатике для проектирования новых белков, молекул, геномов и клеток, а также для изучения основных механизмов и закономерностей биологических систем [4, 5]. Отдельное направление применения генеративного моделирования связано с формированием обучающих данных для методов машинного обучения с целью анализа данных реальных экспериментов [6]. В этом случае формирование смоделированных обучающих данных может иметь преимущества по точности и эффективности при анализе экспериментальных данных, измеренных в различных условиях, обусловленных экспериментальными искажениями. Предполагается, что обуче-

ние на смоделированных данных конкретного эксперимента позволит повысить точность алгоритмов машинного обучения при анализе биофизических систем [7].

В работе предлагается метод генеративного имитационного моделирования для исследования сложных биофизических систем.

ГЕНЕРАТИВНОЕ ИМИТАЦИОННОЕ МОДЕЛИРОВАНИЕ

Предполагается, что через серию экспериментов, порождающих наборы больших данных, исследуется некоторый биофизический объект E, сущностные признаки, или характеристики которого A, должны быть определены в ходе анализа данных. Обозначим через $H_E = \{E_1, E_2, \ldots, E_N\}$ набор данных, состоящий из множества наблюдений, примеров или измерений объекта E, образы или новые проявления которых необходимо сгенерировать. Каждое наблюдение состоит из множества свойств или признаков двух видов: первые (обозначим их X) представляют собой независимые измерения, вторые (обозначим их Y) представляют собой измерения в зависимости от выбранных значений признаков первой группы. Измерения над объектами формируют векторы признаков $X_1, X_2, \ldots, X_K$ и матрицу входных данных X. Цель генеративного моделирования – создать модель, способную генерировать новые наборы признаков ($\dot{X}$, $\dot{Y}$), которые выглядят так, будто созданы с использованием тех же правил, что и исходные данные.

Генеративная модель должна быть вероятностной и включать элементы, которые влияют на отдельные выборки, генерируемые моделью. Можно говорить о том, что существует какое-то неизвестное вероятностное распределение $f(X)$, характеризующее исследуемую систему. Основная задача — создать модель, максимально точно имитирующую распределение $f(X)$, а затем произвести выборку из нее, чтобы сгенерировать новые наблюдения, которые выглядят так, будто могли бы быть в исходном обучающем наборе.

Предполагается, что выполняется анализ больших данных (X, Y), наблюдения H_E экспериментально измерены или сгенерированы в соответствии с некоторым неизвестным распределением $f(X, \Theta) = \{f_1(X_1, \Theta_1), f_2(X_2, \Theta_2), \ldots, f_K(X_K, \Theta_K)\}$, где $f_i(X_i, \Theta_i)$ – функция плотности распределения для признака X_i , Θ_i – вектор параметров распределения f_i. В общем случае решаются задачи *классификации* и *регрессии* [1].

Обобщенный алгоритма генеративного имитационного моделирования.

Шаг 1. Выполняется анализ признаков X с целью выявления законов распределений $f_1(X_1, \Theta_1), f_2(X_2, \Theta_2), \ldots, f_K(X_K, \Theta_K)$ и последующей оценкой

их параметров Θ. Строятся гистограммы h_1, h_2, …, h_K признаков X_1, X_2, …, X_K. Гистограммы аппроксимируются функциями $f'_1(X_1, \Theta'_1)$, $f'_2(X_2, \Theta'_2)$, …, $f'_K(X_K, \Theta'_K)$, приближениями законов распределений $f_1(X_1, \Theta_1)$, $f_2(X_2, \Theta_2)$, …, $f_K(X_K, \Theta_K)$. С использованием метода оптимизации оцениваются параметры Θ'.

Шаг 2. Восстановленные законы распределений $f'(X, \Theta')$ и оценки их параметров Θ' используются в имитационной модели M исследуемых физических характеристик объекта. Имитационная модель генерирует новые данные $(\dot{X}, \dot{Y})$. Правильно подобрав имитационную модель обобщенного алгоритма генеративного имитационного моделирования комплексных биофизических систем, а именно – законы распределений $f'_i(X_i, \Theta'_i)$, $i = 1, 2, \ldots, K$, можно генерировать новые наблюдения, характеризуемые набором данных $(\dot{X}, \dot{Y})$, которые выглядят так, будто были получены из распределения $f(X, \Theta)$.

Шаг 3. Строятся модели машинного обучения $m = \{m_1, m_2, \ldots, m_l\}$ для последующего анализа экспериментальных данных. Смоделированные данные $(\dot{X}, \dot{Y})$ используются для обучения алгоритмов m. Оценивается предсказательная точность моделей машинного обучения.

Шаг 4. Обученные на смоделированных данных алгоритмы машинного обучения m применяются к анализу экспериментальных данных (X, Y). Определяются параметры A.

АДЕКВАТНОСТЬ ГЕНЕРАТИВНЫХ ИМИТАЦИОННЫХ МОДЕЛЕЙ И АНАЛИЗ ТОЧНОСТИ МОДЕЛИРОВАНИЯ

Будем полагать генеративную имитационную модель *адекватной* – если модель с заданной точностью воспроизводит восстановленные законы распределений $f'(X, \Theta')$ и *точной* – если обученные на смоделированных данных модели машинного обучения m оценивают неизвестные параметры A объекта E не хуже или лучше, чем классические алгоритмы анализа данных.

ПРИМЕНЕНИЕ ГЕНЕРАТИВНОГО ИМИТАЦИОННОГО МОДЕЛИРОВАНИЯ К АНАЛИЗУ БИОФИЗИЧЕСКИХ СИСТЕМ

Проверка адекватности разработанных моделей и анализ точности моделирования выполнены на примерах решения задач *классификации* – идентификации сайтов однонуклеотидных полиморфизмов и *регрессии* – предсказания выживаемости онкопациентов на примерах больших наборов данных геномного секвенирования. Полученные результаты позволяют сделать вывод о том, что для реальных экспериментальных данных предпочтительнее использовать модели машинного обучения, обученные

на смоделированных данных. Средняя точность идентификации сайтов однонуклеотидных полиморфизмов на 2-5% выше для моделей машинного обучения, чем для классических статистических методов. При применении генеративного имитационного моделирования для обучения моделей предсказания выживаемости достигнуто почти двукратное снижение ошибки предсказания.

ЗАКЛЮЧЕНИЕ

Разработан метод генеративного имитационного моделирования для исследования сложных биомолекулярных систем по экспериментальным наборам данных, основанный на генерации случайных событий вероятностных моделей распределений, параметры которых оцениваются по имеющимся экспериментальным данным, и методов машинного обучения, обученных на смоделированных данных и применяемых при решении задач *классификации* и *регрессии*. Использование алгоритмов машинного обучения способствует более точному определению закономерностей, получаемых в результате анализа больших экспериментальных данных. Применение разработанного метода позволяет повысить эффективность анализа при идентификации сайтов однонуклеотидных полиморфизмов и предсказании выживаемости онкопациентов по данным геномного секвенирования.

БИБЛИОГРАФИЧЕСКИЕ ССЫЛКИ

1. *Яцков Н. Н., Апанасович В. В.* Комплексный анализ данных при исследовании сложных биомолекулярных систем // Информатика. 2021. Т. 18, № 1. С. 105–122. DOI: https://doi.org/10.37661/1816-0301-2021-18-1-105-122
2. Generative Adversarial Networks and Deep Learning / 1st ed. Eds.: R. Raut, P. D. Pathak, S. R. Sakhare, S. Patil. Boca Raton: Chapman and Hall/CRC, 2023. P. 208.
3. *Foster D.* Generative Deep Learning : Teaching Machines to Paint, Write, Compose, and Play / 2nd ed. Oreilly & Associates Inc., 2023. P. 456.
4. *Ruan X., Murphy R. F.* Evaluation of methods for generative modeling of cell and nuclear shape // Bioinformatics. 2019. Vol. 35(14). P. 2475-2485. DOI: 10.1093/bioinformatics/bty983.
5. *Kucera T., Togninalli M., Meng-Papaxanthos L.* Conditional generative modeling for de novo protein design with hierarchical functions // Bioinformatics. 2022. Vol. 38(13). P. 3454-3461. DOI: 10.1093/bioinformatics/btac353.
6. *Yatskou M.M., Smolyakova E.V., Skakun V.V., Grinev V.V.* Identification of single nucleotide genetic polymorphism sites using machine learning methods // Advances in Transdisciplinary Engineering. 2023. Vol. 42. P. 1031–1037. DOI:10.3233/ATDE231044.
7. *Yatskou M.M., Smolyakova E.V., Skakun V.V., Grinev V.V.* Identification of single nucleotide genetic polymorphism sites using machine learning methods // bioRxiv 2023.10.19.563060. 2023. P. 1–6. DOI: https://doi.org/10.1101/2023.10.19.563060.