УДК 57.087

МЕТРИКА ОЦЕНКИ КОМПОЗИЦИОННОЙ СЛОЖНОСТИ ПЕПТИДОВ ДЛЯ АНАЛИЗА ЭКСТРАПОЛЯЦИОННОЙ СПОСОБНОСТИ КЛАССИФИКАЦИОННЫХ МОДЕЛЕЙ

Дин Е. (ИТМО)

Научный руководитель – кандидат химических наук, Серов Н.С. (ИТМО)

Введение. Производительность на эталонных наборах данных является одним из основных показателей, используемых для оценки современных языковых моделей пептидов (PLM) [1]. Однако в настоящее время отсутствует единый подход к ранжированию эталонных наборов данных по степени сложности: они, как правило, рассматриваются как равнозначные, несмотря на потенциальные различия в структуре и статистических свойствах. Отсутствие формализованной методологии оценки сложности наборов данных затрудняет корректное сравнение моделей и может приводить к неоднозначным выводам об их обобщающей и экстраполяционной способности. В ряде работ активно исследуется проблема сложности наборов данных для табличных и текстовых данных, однако аналогичные исследования практически отсутствуют для биологических последовательностей [2–3]. В связи с этим возникает необходимость разработки единой скалярной метрики композиционной сложности, основанной на свойствах первичной структуры пептидов, а также учитывающей характеристики самой задачи классификации, включая линейную разделимость классов и различия между распределениями выборок, оцениваемые с использованием методов оптимального транспорта. Такая метрика позволит ранжировать существующие эталонные наборы данных и обеспечит более корректное и интерпретируемое сравнение PLM, выявляя их сильные и слабые стороны с точки зрения экстраполяционной способности.

Основная часть. В данной работе исследование сосредоточено на задачах классификации, поскольку именно этот тип задач наиболее широко используется при сравнении моделей на стандартных бенчмарках. В данном исследовании были использованы сведения о более 50 эталонных наборов данных для тестирования производительности языковых моделей пептидов, которые покрывают задачи бинарной классификации связанные с антимикробной, противовирусной и противоопухолевой активностью пептидов, а также задачи классификации клеточно-проникающих пептидов и прогнозирования лекарственной устойчивости ВИЧ [4]. В них содержится информация исключительно о первичной структуре пептида и целевом классе. Основной целью было выделить корреляцию характеристик сложности набора данных с производительностью моделей. Для этого мы обучили алгоритмы машинного обучения разных категорий, которые включают обобщенные линейные модели (логистическая регрессия), методы, основанные на маргинальных переменных (метод опорных векторов), непараметрические подходы (метод k-ближайших соседей) и ансамблевые алгоритмы (случайные леса и XGBoost). Для каждой модели были зафиксированы средние значения производительности на 5-кратной стратифицированной перекрестной проверке. Затем были рассчитаны статистические показатели, объединенные в три логические группы характеристик: параметры, отражающие разнообразие последовательностей, меры сложности задачи классификации, а также показатель оптимального транспорта между двумя классами. Далее была поставлена задача выявления корреляций отдельных характеристик, а также их комбинаций, с метриками производительности моделей [5]. Для этого анализировались коэффициенты корреляции Пирсона и Спирмена. Дополнительно были построены интерпретируемые модели, включая линейную регрессию и дерево принятия решений, что позволило выявить совместное влияние всех характеристик и оценить вклад каждой из них.

Выводы. Таким образом, разработка унифицированной метрики сложности наборов данных позволит выстроить иерархическую систему оценки результативности существующих языковых моделей. Предлагаемая метрика обеспечит ранжирование бенчмарков и проведение компаративного анализа моделей, устанавливая взаимосвязь между их производительностью и внутренними характеристиками данных. Особую ценность представляет применение метрики на этапе агрегации выборок: она идентифицирует аспекты, обусловливающие остаточную неразделимость задач, что может быть скорректировано за счет целевого пополнения обучающих примеров. Важно отметить, что предложенный подход масштабируем и применим к широкому кругу биологических объектов, для которых первичная структура является определяющим источником информации, включая белки, ДНК и РНК, что будет изучаться на следующем этапе работы.

Список использованных источников:

1. Ankh: Optimized protein language model unlocks general-purpose modelling / Elnaggar, A., Essam, H., Salah-Eldin, W. [et al.]. – DOI 10.48550/arXiv.2301.06568 // arXiv preprint. — 2023. — arXiv:2301.06568. – URL: https://doi.org/10.48550/arXiv.2301.06568.
2. How complex is your classification problem? A survey on measuring classification complexity / Lorena, A. C., Garcia, L. P., Lehmann, J. [et al.]. – DOI 10.1145/3347711 // ACM Computing Surveys (CSUR). — 2019. — Vol. 52, No 5. – p. 1-34. – URL: https://doi.org/10.1145/3347711.
3. Evolutionary data measures: Understanding the difficulty of text classification tasks / Collins, E., Rozanov, N., Zhang, B. – DOI 10.48550/arXiv.1811.01910 // arXiv preprint. — 2018. — arXiv:1811.01910. – URL: https://doi.org/10.48550/arXiv.1811.01910.
4. A large-scale comparative study on peptide encodings for biomedical classification / Spänig, S., Mohsen, S., Hattab, G. [et al.]. – DOI 10.1093/nargab/lqab039 // NAR Genomics and Bioinformatics. — 2021. — Vol. 3, No 2. – lqab039. – URL: https://doi.org/10.1093/nargab/lqab039.
5. Choosing the Right Dataset: Hardness Criteria for Feature Selection Benchmarking / Elmakias, I., Vilenchik, D. – DOI 10.1016/j.knosys.2025.115022 // Knowledge-Based Systems. — 2025. — Vol. 312. – 115022. – URL: https://doi.org/10.1016/j.knosys.2025.115022.