НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
июль–август 2024 Том 24 № 4 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
July– August 2024 Vol. 24 No 4 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

doi: 10.17586/2226-1494-2024-24-4-594-601
УДК 004.855.5: 004.032.26

Предсказание связей «ген-болезнь» с помощью гетерогенной графовой нейронной сети

Денис Александрович Сидоренко[1]✉, Анатолий Абрамович Шалыто[2]

[1,2] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация

[1] denisissveta@mail.ru✉, https://orcid.org/0009-0004-0571-5192
[2] shalyto@mail.ifmo.ru, https://orcid.org/0000-0002-2723-2077

Аннотация

Введение. Представлены результаты разработки модели гетерогенной графовой нейронной сети для предсказания ассоциаций между генами и заболеваниями на основе имеющихся геномных и медицинских данных. Новизна предложенного подхода состоит в объединении концепций графовых нейронных сетей и гетерогенных информационных сетей для эффективной обработки структурированных данных и учета сложных взаимодействий между генами и патологиями. **Метод.** Предложенное решение представляет собой гетерогенную графовую нейронную сеть, которая использует гетерогенную графовую структуру для представления генов, болезней и их взаимосвязей. **Основные результаты.** Оценка точности разработанной модели проведена на наборах данных DisGeNET, LASTFM, YELP. На этих же данных выполнено сравнение результатов с наиболее сильными моделями. Показано превосходство предложенной модели по метрикам точности Average Precision (AP), F1-меры (F1@S), Hit@k, Area Under Receiver Operating Characteristic curve (AUROC) при предсказании ассоциаций «ген-болезнь». **Обсуждение.** Разработанная модель может использоваться как инструмент биоинформатического анализа и в качестве вспомогательного средства для исследователей и врачей при изучении генетических заболеваний. Такой подход может ускорить процесс открытия новых лекарственных мишеней и разработку персонализированной медицины.

Ключевые слова

машинное обучение, графовые нейронные сети, гетерогенные информационные сети, биоинформатика, генетика, предсказание «ген-болезнь» ассоциаций

Ссылка для цитирования: Сидоренко Д.А., Шалыто А.А. Предсказание связей «ген-болезнь» с помощью гетерогенной графовой нейронной сети // Научно-технический вестник информационных технологий, механики и оптики. 2024. Т. 24, № 4. С. 594–601. doi: 10.17586/2226-1494-2024-24-4-594-601

Predicting gene-disease associations using a heterogeneous graph neural network

Denis A. Sidorenko[1]✉, Anatoly A. Shalyto[2]

[1,2] ITMO University, Saint Petersburg, 197101, Russian Federation

[1] denisissveta@mail.ru✉, https://orcid.org/0009-0004-0571-5192
[2] shalyto@mail.ifmo.ru, https://orcid.org/0000-0002-2723-2077

Abstract

The research presents the development of a heterogeneous graph neural network model for predicting gene-disease using existing genomic and medical data. The novelty of the approach is in integrating the principles of graph neural networks and heterogeneous information networks for efficient processing of structured data and consideration of complex gene-pathology interactions. The solution proposed is a heterogeneous graph neural network which utilizes a heterogeneous graph structure for representing genes, diseases, and their relationships. The performance of the developed model was evaluated on the DisGeNET, LASTFM, YELP datasets. On these datasets, a comparison was made with current SOTA models. The comparison results demonstrated that the proposed model outperforms other models in terms of Average Precision (AP), F1-measure (F1@S), Hit@k, Area Under Receiver Operating Characteristic curve (AUROC)

in predicting “gene-disease” associations. The model developed serves as a tool for bioinformatics analysis and can aid researchers and doctors in studying genetic diseases. This could expedite the discovery of new drug targets and the advancement of personalized medicine.

Keywords
machine learning, graph neural networks, heterogeneous information networks, bioinformatics, genetics, “gene-disease” prediction associations

For citation: Sidorenko D.A., Shalyto A.A. Predicting gene-disease associations using a heterogeneous graph neural network. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2024, vol. 24, no. 4, pp. 594–601 (in Russian). doi: 10.17586/2226-1494-2024-24-4-594-601

Введение

Гетерогенные графовые нейронные сети стали важнейшим инструментом в области искусственного интеллекта и машинного обучения, особенно при работе со сложными структурами данных. В контексте графовых нейронных сетей гетерогенные графы относятся к графам, которые содержат вершины и ребра разных типов, что обеспечивает более полное представление реальных данных. В отличие от традиционных графовых нейронных сетей [1], которые ориентированы на однородные графы, гетерогенные графовые нейронные сети предназначены для обработки различных типов отношений данных, таких как взаимодействие пользователя с элементами, ассоциации атрибутов и отношения контента [2]. Объединив это разнообразие, гетерогенные графовые нейронные сети могут отражать более сложные закономерности и зависимости в данных, что приводит к повышению производительности при решении различных задач. Приложения таких моделей охватывают широкий спектр областей, включая системы социальных рекомендаций, платформы электронной коммерции, биоинформатику и графы знаний [3]. В последнее время развиваются модели по поиску связей между генами и болезнями, основанные на решении задачи Link Prediction на графах с помощью распространения сигнала по графу на основе случайного блуждания по нему [4]. В частности, для решения этой задачи с использованием случайных блужданий были разработаны такие модели как PRINCE и ORIENT [5].

В модели PRINCE построена приоритезация генов с помощью методов Random Walks with Restart и присвоены априорные вероятности генам исходя из близости болезней, а в ORIENT — дополнительно усилены априорные вероятности через кратчайшие пути от гена до болезни. Дополнительно для решения задачи Link Prediction в работе [6] разработана сетевая модель GuityTargets, которая до появления графовых нейронных сетей являлась наиболее эффективной применительно к этой задаче.

Основной недостаток моделей, основанных на случайном блуждании или сетевом подходе, состоит, в том, что информация плохо распространяется по графу, они неэффективны на больших и сложных графах. Графовые нейронные сети позволяют создавать более мощные модели, которые охватывают как структурную, так и контекстную информации. В [7] создана модель Progressive Graph Convolution Network, которая реализует графовые свертки. Однако она обладает существенными недостатками: в ней архитектурно используются моногенные свертки на графе — в модели вершины разных типов обозначают то же самое; не были использованы последние достижения из сложных трансформерных графовых сверток.

В настоящей работе предложена модель Heterogeneous Graph Gene Disease Link Prediction (HeteroGGDLP), которая объединяет в себе, как распространение гетерогенного сигнала по графу, так и элементы новейших архитектур графовых нейронных сетей. Кроме того, в модели применен графовый трансформер для предсказания связей «ген-болезнь», а также уникальная инициализация, которая в этой задаче ранее не применялась. В модель, используя механизм внимания, введен адаптивный механизм агрегации соседей. При агрегации информации (атрибутами и структурой окружения вершины) между вершинами применяется иерархический подход.

Использование гетерогенных графов в области нейронных сетей позволило разрешить сложные противоречия между различными объектами. Эти графы состоят из различных вершин, которые символизируют такие сущности, как отдельные гены или болезни, и ребер, обозначающих связи между этими сущностями. В отличие от однородных графов, в гетерогенных графах используются несколько типов вершин и ребер, позволяющих распознавать сложные закономерности и корреляции между различными типами данных. Нейронные сети с гетерогенными графами [8] используют такую структуру для выявления связей между разнородными типами данных, что делает их неоценимыми для выявления сложных взаимозависимостей данных. Учитывая гетерогенный (имеющий разную природу отдельных вершин графа) характер информации в графах, графовые нейронные сети обычно используют слои эмбеддингов (функций трансформаций) для преобразования дискретных данных (такие как вершины и ребра графа) в векторные представления с вещественными значениями, что способствует эффективной обработке данных, так как появляется механизм для обработки структурированных данных графа с помощью стандартных операций, применяемых к непрерывным векторам. Специализированные свертки в графовых нейронных сетях, такие как Relational Graph Convolutional Networks [9], предназначены для работы с различными типами ребер в гетерогенных графах, тем самым обучая отдельные матрицы для обработки разнообразных связей внутри графа. Также известны модели, которые используют механизм внимания в своих архитектурах для различных типов ребер [10].

Метод

Формулировка математической проблемы. Сформулируем задачу Link Prediction (предсказание связей) на графе, где гены представлены как белок-белковые взаимодействия (Protein-Protein Interactions, PPI), а болезни — как онтология экспериментальных факторов (Experimental Factor Ontology, EFO).

Пусть задан гетерогенный граф $G = (V, E)$, где V — множество вершин, представляющих гены (PPI [11]) и болезни (EFO [12]); E — множество ребер, представляющих связи между генами и болезнями. Более формально: $V = V_{PPI} \cup V_{EFO}$ — объединение вершин из двух наборов данных. Ребра, $E = E_{gg} \cup E_{gd} \cup E_{dd}$ соответственно, можно разделить на три типа: между генами, между генами и болезнями, между болезнями.

Таким образом, граф G содержит информацию о взаимодействиях белков, представленных генами, а также об ассоциациях генов с болезнями. Задача Link Prediction состоит в предсказании отсутствующих или новых ребер E_{gd} между генами и болезнями, используя структуру графа, а также информацию, содержащуюся в существующих ребрах.

Формально эта задача может быть сформулирована как бинарная классификационная задача, в которой для каждой пары (u, v), где $u \in V_{PPI}$, а $v \in V_{EFO}$, необходимо предсказать, существует ли ребро $(u, v) \in E_{gd}$. Цель исследования — обучить модель, способную предсказывать новые ассоциации между генами и болезнями, на основе информации о взаимодействиях белков и известных ассоциациях «ген-болезнь», представленных в графе G.

Разработка и построение модели. Модель HeteroGGDLP, основанная на базе (Heterogeneous Graph Transformer)-сверток (HGT) на (PPI, EFO)-графе, расширяет классическую архитектуру за счет применения следующих этапов.

1. Неслучайная инициализация — сеть, инициализируемая предварительной текстовой информацией, эмбеддинги которой получены с помощью модели BioBERT [13]. Пусть t_v — текстовые описания вершины v, тогда $e_v = BioBERT(t_v)$ — инициализация векторного представления вершины.
2. Адаптивный механизм агрегации соседей. Пусть N_v — множество соседей вершины v; $h_v^{(l)}$ — представление вершины в слое l; $\mathbf{W}_{self}^{(l)}$ — матрица весов для представления вершины в слое l; $\mathbf{W}_{neigh}^{(l)}$ — матрица весов для представления *neigh* соседей в слое l. Вместо стандартной агрегации соседей по формуле:

$$h_v^{(l+1)} = \sigma\left(\mathbf{W}_{self}^{(l)} h_v^{(l)} + \sum_{u \in N_v} \mathbf{W}_{neigh}^{(l)} h_u^{(l)}\right),$$

используем механизм на основе внимания:

$$h_v^{(l+1)} = \sigma\left(\mathbf{W}_{self}^{(l)} h_v^{(l)} + \sum_{u \in N_v} \alpha_{vu}^{(l)} \mathbf{W}_{neigh}^{(l)} h_u^{(l)}\right),$$

где $\alpha_{vu}^{(l)} = \dfrac{\exp\left(score(h_v^{(l)}, h_u^{(l)})\right)}{\sum_{k \in N_v}\left(score(h_v^{(l)}, h_u^{(l)})\right)}$, $score\left(\left(h_v^{(l)}, h_u^{(l)}\right)\right)$ — функция, определяющая внимание. Это, как правило, LeakyRelu [14].

3. Иерархическая структура агрегации. Вместо плоской агрегации соседей в модель введем иерархическую структуру, где сначала агрегируются соседи нижнего уровня:

$$h_{v,j}^{(l+1)} = \sigma\left(\mathbf{W}_{self}^{(l)} h_{v,j}^{(l)} + \sum_{u \in N_{v,j}} \mathbf{W}_{neigh}^{(l)} h_u^{(l)}\right),$$

где $N_{v,j}$ — множество соседей нижнего уровня j вершины v; $h_{v,j}^{(l)}$ — представление нижнего уровня j вершины v в слое l. Затем представления соседей нижнего уровня агрегируются для получения представления вершин верхнего уровня:

$$h_v^{(l+1)} = \sigma\left(\mathbf{W}_{self}^{(l)} h_v^{(l)} + \sum_j \mathbf{W}_{high}^{(l)} h_{v,j}^{(l+1)}\right).$$

4. Таким образом, объединяя адаптивный механизм агрегации соседей и иерархическую структуру агрегации, получим формулы, для агрегации представлений (эмбеддингов), объединяющие эти подходы: $h_{v,j}^{(l+1)} = \sigma\left(\mathbf{W}_{self}^{(l)} h_{v,j}^{(l)} + \sum_{u \in N_{v,j}} \alpha_{vu}^{(l)} \mathbf{W}_{neigh}^{(l)} h_u^{(l)}\right)$ — для нижнего и $h_v^{(l+1)} = \sigma\left(\mathbf{W}_{self}^{(l)} h_v^{(l)} + \sum_j \beta_{vj}^{(l)} \mathbf{W}_{high}^{(l)} h_{v,j}^{(l+1)}\right)$ — для верхнего уровней, где $\beta_{vj}^{(l)}$ — вес внимания для представления соседа нижнего уровня j вершины v в слое l, вычислим по формуле:

$$\beta_{vj}^{(l)} = \frac{\exp\left(score(h_v^{(l)}, h_{v,j}^{(l)})\right)}{\sum_{k \in N_v}\left(score(h_v^{(l)}, h_{v,k}^{(l)})\right)}.$$

Механизм внимания используем на нижнем уровне для агрегации представлений соседей в представления соседей нижнего уровня и на верхнем уровне для агрегации представлений соседей нижнего уровня в итоговое представление вершины. Данный подход позволяет модели адаптивно агрегировать информацию с учетом структуры графа и важности отдельных соседей на разных уровнях для различных типов связей.

Модель обучается, изменяя матрицы весов $\mathbf{W}_{neigh}^{(l)}$ через агрегацию представлений соседей различных типов связей на нижнем уровне, а затем объединяет эту информацию на верхнем уровне с учетом важности каждого типа.

В целом архитектура модели представляет собой несколько компонент:

1) слой инициализации BioBERT эмбеддингами;
2) энкодер, внутри которого введена иерархическая агрегация вершин, веса внимания (трансформерная часть) для определения важности вершины в момент агрегации, а также матрица весов, изменяющихся во время обучения модели. Веса модели обучаются на разных уровнях модели: в графовых свертках, на каждом уровне иерархии агрегации;
3) декодер, который представляет собой функцию вычисления близости вершин и функцию активации для определения вероятности наличия связи. Выход декодера отправляется в функцию потерь, после расчета которой — вычисляется потеря. В процессе обучения модели для каждой пары вершин (u, v) декодер вычисляет оценку вероятности наличия

ребра между ними на основе сходства их векторных представлений:

$$p_{uv} = DEC(h_u, h_v) = \sigma(h_u^T h_v).$$

Функция потерь — бинарная кросс-энтропия, которая определяется по формуле:

$$L = \sum_{(u,v)\in E} \log(p_{uv}) + (1 - y_{uv})\log(1 - p_{uv}),$$

где $y_{uv} = 1$, если $(u, v) \in E$ и $y_{uv} = 0$ в противном случае. Оптимизация параметров энкодера и декодера производится путем минимизации функции потерь L методом стохастического градиентного спуска, во время которого вычисляются градиенты и обновляются веса во всей модели для минимизации функции потерь. После обучения на выходе энкодера формируются обученные векторные представления вершин, а на выходе декодера — вероятности связей, позволяющие оценивать качество модели на любых классификационных метриках.

На рисунке показана архитектура представленной модели HeteroGGDLP, которая содержит блок BioBERT для преобразования входных данных, а также блоки энкодера и декодера.

Основные результаты

Наборы данных. Для оценки качества и сравнения предложенной модели с существующими моделями были использованы три открытых набора данных с различными характеристиками:

— DisGeNET [15] — ключевой набор данных для задачи предсказания ассоциаций «ген-болезнь»;
— YELP [16] — набор данных, содержащий информацию о связях между пользователями, бизнесами и возможными действиями между ними;
— LASTFM [17] — набор данных, описывающий взаимодействия между пользователями, музыкантами/группами и тегами.

Задача Link Prediction на наборах данных (табл. 1) сформулирована как бинарная задача классификации. Для каждого набора данных случайным образом была удалена часть связей для формирования тестового набора. Включение разнообразных наборов данных, таких как YELP и LASTFM, кроме DisGeNET, позволяет продемонстрировать широкую применимость предложенной модели и оценить ее производительность в различных контекстах предсказания связей на гетерогенных графах, не ограничиваясь только задачей предсказания ассоциаций «ген-болезнь».

Модели для сравнения. Для сравнения с предложенной моделью HeteroGGDLP выбраны наиболее сильные модели для предсказания связей на гетерогенных графах.

— HerGePred [18] — модель для предсказания связей в гетерогенных сетях, основанная на случайных блужданиях, использующая модель node2vec [19], для того, чтобы генерировать конкретные вероятности перехода в зависимости от шага в глубину или в ширину.
— Metapath2Vec [20] — обобщение популярного алгоритма word2vec [21] на графах, позволяющее использовать векторные представления в гетерогенном графе с помощью техники Random Walks [4] для метапутей, где метапуть — заранее заданная последовательность вершин различных типов.
— DeepWalk [22] — классическая модель обучения векторных представлений вершин в однородных графах на основе техники Random Walks и word2vec. Не учитывает особенности вершин/ребер.
— HGT [23] — одна из новейших моделей трансформеров, способная обрабатывать разнородные графы с учетом использования различных типов вершин и механизма внимания.
— Heterogeneous Graph Attention Network (HAN) [10] — модель на основе графовых сверточных сетей, в которой применяется специализированный иерархический механизм для агрегации информации из метапутей в гетерогенном графе.

Таблица 1. Статистики по наборам данных для тестирования

Table 1. Statistics for testing datasets

Набор данных	Число вершин	Число ребер	Число типов ребер
DisGeNET	56 379	587 775	4
LASTFM	20 612	128 804	3
YELP	94 807	1 406 809	9

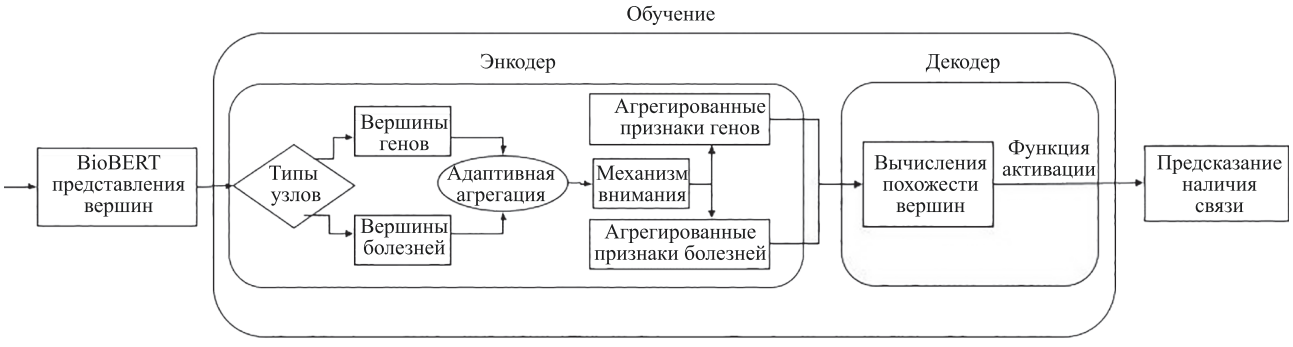

Рисунок. Архитектура модели HeteroGGDLP

Figure. Architecture of the HeteroGGDLP model

— Factor Heterogeneous Network Embedding (FHNE) [24] — модель, которая объединяет метапутевой механизм и механизм факторизации семантической информации в единую архитектуру глубокого обучения для извлечения представлений из гетерогенных графов.

Метрики

Для расчета метрик бинарной классификации исходный граф был разбит на тренировочный и тестовый. Для каждого ребра из тестового графа модель HeteroGGDLP предсказала вероятность принадлежности к положительному классу (наличие ребра). Затем предсказанные вероятности сортировались в порядке убывания. Полученный упорядоченный список вероятностей был использован для следующих метрик.

1. Area Under the Receiver Operating Characteristic Curve (AUROC) — метрика, показывающая способность модели ранжировать положительные (есть ребро) и отрицательные (нет ребра) примеры. Значение AUROC варьируется от 0 до 1, где 1 — идеальный результат.
2. Average Precision (AP) — усредненная точность. Чем выше значение метрики, тем больше модель ранжирует положительные примеры.
3. Hit@k — доля верных предсказаний в топ-k результатах. Чем выше значение метрики, тем больше релевантных связей модель предсказывает в первых k результатах (сколько правильных предсказаний получается в топ-k ранжированном списке связей).
4. F1@S — F1-мера (гармоническое среднее между точностью и полнотой) для топ-S результатов. Показывает общую производительность на топ-S предсказаниях.

Анализ экспериментальных данных

В табл. 2 представлены результаты модели HeteroGGDLP по предсказанию связей на тестовых наборах данных.

В результате анализа табл. 2, можно сделать вывод, что предложенная модель HeteroGGDLP конкурен-

Таблица 2. Сравнение моделей в наборах данных DisGeNET, LASTFM и YELP
Table 2. Comparison of models on the DisGeNET, LASTFM and YELP datasetes

Модели	Метрики					
	AUROC (±std)	AP (±std)	Hit@10 (±std)	Hit@100 (±std)	Hit@S (±std)	F1@S (±std)
DisGeNET						
HerGePred	0,490 ± 0,421	0,700 ± 0,231	0,235 ± 0,180	**0,029 ± 0,031**	0,588 ± 0,041	0,490 ± 0,414
Metapath2Vec	0,577 ± 0,450	0,775 ± 0,229	**0,241 ± 0,175**	**0,029 ± 0,031**	**0,632 ± 0,051**	0,565 ± 0,443
DeepWalk	0,722 ± 0,387	0,837 ± 0,207	0,207 ± 0,156	0,025 ± 0,027	0,532 ± 0,033	0,702 ± 0,385
HGT	**0,873 ± 0,292**	**0,922 ± 0,159**	0,196 ± 0,160	0,024 ± 0,027	0,482 ± 0,067	**0,858 ± 0,297**
HAN	0,767 ± 0,314	0,817 ± 0,218	0,190 ± 0,172	0,023 ± 0,029	0,437 ± 0,202	0,772 ± 0,361
FHNE	**0,868 ± 0,294**	**0,917 ± 0,161**	0,206 ± 0,158	0,025 ± 0,027	0,543 ± 0,037	**0,852 ± 0,298**
HeteroGGDLP	0,836 ± 0,331	0,908 ± 0,173	**0,240 ± 0,181**	**0,029 ± 0,031**	**0,618 ± 0,039**	**0,818 ± 0,334**
LASTFM						
HerGePred	0,543 ± 0,138	0,591 ± 0,114	**0,605 ± 0,040**	0,188 ± 0,021	**0,599 ± 0,011**	0,522 ± 0,118
Metapath2Vec	0,625 ± 0,168	0,687 ± 0,137	0,602 ± 0,039	**0,190 ± 0,021**	0,598 ± 0,011	0,584 ± 0,151
DeepWalk	0,393 ± 0,231	0,504 ± 0,156	0,520 ± 0,037	0,159 ± 0,018	0,512 ± 0,018	0,414 ± 0,190
HGT	**0,790 ± 0,180**	**0,830 ± 0,142**	0,514 ± 0,034	0,154 ± 0,018	0,513 ± 0,007	**0,752 ± 0,164**
HAN	0,750 ± 0,157	0,778 ± 0,123	0,597 ± 0,049	0,154 ± 0,024	0,579 ± 0,025	0,710 ± 0,138
FHNE	0,768 ± 0,153	0,789 ± 0,131	0,561 ± 0,049	0,161 ± 0,019	0,544 ± 0,030	0,724 ± 0,141
HeteroGGDLP	**0,847 ± 0,115**	**0,857 ± 0,107**	**0,655 ± 0,049**	**0,200 ± 0,023**	**0,647 ± 0,026**	**0,783 ± 0,119**
YELP						
HerGePred	0,498 ± 0,256	0,638 ± 0,173	**0,416 ± 0,117**	**0,042 ± 0,012**	0,620 ± 0,017	0,498 ± 0,227
Metapath2Vec	**0,865 ± 0,179**	**0,892 ± 0,140**	0,404 ± 0,114	0,041 ± 0,012	0,607 ± 0,018	**0,816 ± 0,199**
DeepWalk	0,604 ± 0,273	0,708 ± 0,192	0,338 ± 0,095	0,034 ± 0,010	0,503 ± 0,017	0,584 ± 0,246
HGT	0,707 ± 0,241	0,780 ± 0,174	0,347 ± 0,101	0,035 ± 0,011	0,523 ± 0,017	0,665 ± 0,227
HAN	0,578 ± 0,247	0,691 ± 0,173	0,349 ± 0,097	0,035 ± 0,010	0,502 ± 0,001	0,561 ± 0,225
FHNE	0,656 ± 0,248	0,749 ± 0,177	0,411 ± 0,134	0,041 ± 0,014	**0,669 ± 0,067**	0,624 ± 0,228
HeteroGGDLP	**0,927 ± 0,132**	**0,940 ± 0,108**	**0,414 ± 0,117**	**0,042 ± 0,012**	**0,629 ± 0,020**	**0,887 ± 0,167**

Примечание. (±std) — стандартное отклонение.

тоспособна на наборе данных DisGeNET с другими современными моделями. Несмотря на то, что по отдельным метрикам, таким как AUROC и AP, модели FHNE и HGT показывают более высокие результаты, HeteroGGDLP обеспечивает лучшие значения по другим метрикам, включая Hit@10 (0,240) и F1@S (0,818).

Хотя метрики AUROC и AP отражают общую способность модели HeteroGGDLP правильно ранжировать положительные и отрицательные примеры, метрики Hit@k и F1@S более точно измеряют практическую эффективность модели в выявлении релевантных связей «ген-болезнь». В этом контексте преимущество модели HeteroGGDLP по метрикам Hit@10 и F1@S указывает на ее способность успешно извлекать верные ассоциации «ген-болезнь» по сравнению с другими моделями (значения выделены жирным шрифтом в табл. 2), что имеет значение для практического применения в биомедицинских исследованиях, и поэтому модель HeteroGGDLP демонстрирует лучшее качество на наборе данных DisGeNET.

На наборе данных LastFM предложенная модель HeteroGGDLP продемонстрировала высокие результаты по всем ключевым метрикам, опережая другие рассматриваемые модели. В частности, HeteroGGDLP показала лучшие значения метрик AUROC (0,847) и AP (0,857), что свидетельствует о высокой способности этой модели правильно ранжировать положительные и отрицательные примеры. Кроме того, модель HeteroGGDLP превзошла остальные модели по значениям метрик Hit@10 (0,655), Hit@100 (0,200) и F1@S (0,783), что указывает на ее эффективность в выявлении верных ассоциаций из верхней части списка. Несмотря на то, что модели HGT и HAN также продемонстрировали конкурентоспособные результаты по таким метрикам, как AUROC и AP, их качество по метрикам Hit@k и F1@S заметно уступает модели HeteroGGDLP.

На наборе данных YELP модель HeteroGGDLP показала наилучшие значения метрик AUROC (0,927) и AP (0,940), что указывает на способность корректно ранжировать положительные и отрицательные примеры. Следует отметить, что по этим же метрикам модель Metapath2Vec показала второй результат после HeteroGGDLP со значениями AUROC (0,865) и AP (0,892). Преимуществом модели HeteroGGDLP является ее эффективность в извлечении значимых релевантных связей из верхней части ранжированного списка. Это видно по значениям Hit@10 (0,414), Hit@100 (0,042) и F1@S (0,887). Несмотря на схожие результаты остальных моделей по отдельным метрикам, в целом модель HeteroGGDLP продемонстрировала лучший результат по совокупности метрик.

Обсуждение

Результаты экспериментов на различных наборах данных, представленных в табл. 2, продемонстрировали высокую эффективность предложенной модели HeteroGGDLP в задачах извлечения и прогнозирования связей в сложных гетерогенных графовых структурах. Наиболее высокие результаты были достигнуты на наборе данных YELP, где модель HeteroGGDLP показала лучшие значения по таким критически важным метрикам, как AUROC, AP, Hit@10, Hit@100 и F1@S, опережая альтернативные модели.

Вместе с тем, на других наборах данных, таких как DisGeNET и LASTFM, некоторые современные модели (FHNE и HGT), которые учитывают гетерогенность графов, показали схожие или более высокие результаты по отдельным метрикам. Это подчеркивает сложность задачи извлечения связей из разнородных графовых данных и необходимость дальнейших исследований для повышения стабильности и универсальности предлагаемых подходов на различных типах графовых структур.

Тем не менее, предложенная модель HeteroGGDLP продемонстрировала высокий результат по совокупности значений различных метрик, применяемых для оценки моделей на тестовых наборах данных. Помимо HeteroGGDLP, другие современные модели (HGT, HAN и FHNE) также показали хорошие результаты в сравнении с базовыми моделями (DeepWalk, Metapath2Vec), изначально разработанными для гомогенных графов. Это демонстрирует важность учета разнородности типов вершин и ребер в гетерогенных моделях для повышения качества извлечения и прогнозирования связей.

Заключение

Представлено описание разработанной модели HeteroGGDLP — гетерогенной графовой нейронной сети для задачи прогнозирования ассоциаций между генами и заболеваниями. Модель основана на совместном использовании механизмов трансформеров и графовых нейронных сетей для эффективной обработки разнородных графовых структур, представляющих взаимосвязи между генами и заболеваниями.

Достигнутые результаты имеют важное практическое значение для биомедицинских приложений, таких как поиск новых терапевтических мишеней, диагностика заболеваний и разработка персонализированной медицины.

Отметим ряд ограничений текущей модели. Во-первых, использованные наборы данных могут быть расширены путем интеграции дополнительных источников биомедицинской информации. Во-вторых, необходимо дальнейшее совершенствование архитектуры модели и методов обучения для повышения стабильности и способности к обобщению на разнообразных типах графовых структур. В-третьих, требуется учет дополнительных факторов и контекстной информации, влияющих на ассоциации «ген-болезнь», таких как экспрессия генов, эпигенетические модификации, клинические и демографические данные пациентов.

Литература

1. Henaff M., Bruna J., LeCun Y. Deep convolutional networks on graph-structured data // arXiv. 2015. arXiv:1506.05163. https://doi.org/10.48550/arXiv.1506.05163
2. Wang X., Bo D., Shi C., Fan S., Ye Y., Yu P.S. A survey on heterogeneous graph embedding: methods, techniques, applications and sources // IEEE Transactions on Big Data. 2023. V. 9. N 2. P. 415–436. https://doi.org/10.1109/TBDATA.2022.3177455
3. Shao B., Li X., Bian G. A survey of research hotspots and frontier trends of recommendation systems from the perspective of knowledge graph // Expert Systems with Applications. 2021. V. 165. P. 113764. https://doi.org/10.1016/j.eswa.2020.113764
4. László L. Random walks on graphs: a survey // Combinatorics. V. 2. 1993. P. 1–46.
5. Li L., Wang Y., An L., Kong X., Huang T. A network-based method using a random walk with restart algorithm and screening tests to identify novel genes associated with Menière's disease // PLOS ONE. 2017. V. 12. N 8. P. e0182592. https://doi.org/10.1371/journal.pone.0182592
6. Muslu Ö., Hoyt C.T., Lacerda M., Hofmann-Apitius M., Frohlich H. GuiltyTargets: Prioritization of novel therapeutic targets with network representation learning // IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2022. V. 19. N 1. P. 491–500. https://doi.org/10.1109/TCBB.2020.3003830
7. Li Y., Kuwahara H., Yang P., Song L., Gao X. PGCN: Disease gene prioritization by disease and gene embedding through graph convolutional neural networks // biorxiv.org. 2019. https://doi.org/10.1101/532226
8. Dutta A., Alcaraz J., TehraniJamsaz A., Cesar E., Sikora A., Jannesari A. Performance optimization using multimodal modeling and heterogeneous GNN // arXiv. 2023. arXiv.2304.12568. https://doi.org/10.48550/arXiv.2304.12568
9. Thanapalasingam T., van Berkel L., Bloem P., Groth P. Relational graph convolutional networks: Closer Look // PeerJ Computer Science. 2022. V. 8. P. e1073. https://doi.org/10.7717/PEERJ-CS.1073
10. Wang X., Ji H., Shi C., Wang B., Ye Y., Cui P., Yu P.S. Heterogeneous graph attention network // Proc. of the WWW '19: The World Wide Web Conference. 2019. P. 2022–2032. https://doi.org/10.1145/3308558.3313562
11. Ali A., Bagchi A. An overview of protein-protein interaction // Current Chemical Biology. 2015. V. 9. N 1. P. 53–65. https://doi.org/10.2174/2212796809011511 09161126
12. Malone J., Holloway E., Adamusiak T., Kapushesky M., Zheng J., Kolesnikov N., Zhukova A., Brazma A., Parkinson H. Modeling sample variables with an experimental factor ontology // Bioinformatics. 2010. V. 26. N 8. P. 1112–1118. https://doi.org/10.1093/bioinformatics/btq099
13. Lee J., Yoon W., Kim S., Kim D., Kim S., So C.H., Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining // Bioinformatics. 2020. V. 36. N 4. P. 1234–1240. https://doi.org/10.1093/bioinformatics/btz682
14. Zhang X., Zou Y., Shi W. Dilated convolution neural network with LeakyReLU for environmental sound classification // Proc. of the 22nd International Conference on Digital Signal Processing (DSP). 2017. https://doi.org/10.1109/ICDSP.2017.8096153
15. Piñero J., Queralt-Rosinach N., Bravo A., Deu-Pons J., Bauer-Mehren A., Baron M., Sanz F., Furlong L.I. DisGeNET: A discovery platform for the dynamical exploration of human diseases and their genes // Database. 2015. V. 2015. https://doi.org/10.1093/database/bav028
16. Alam M., Cevallos B., Flores O., Lunetto R., Yayoshi K., Woo J. Yelp Dataset Analysis using Scalable Big Data // arXiv. 2021. arXiv.2104.08396v1. https://doi.org/10.48550/arXiv.2104.08396
17. Li Y., Guo X., Lin W., Zhong M., Li Q., Liu Z., Zhong W., Zhu Z. Learning dynamic user interest sequence in knowledge graphs for click-through rate prediction // IEEE Transactions on Knowledge and Data Engineering. 2023. V. 35. N 1. P. 647–657. https://doi.org/10.1109/TKDE.2021.3073717
18. Kuo Y., Wang R., Liu G., Shu Z., Wang N., Zhang R., Yu J., Chen J., Li X., Zhou X. HerGePred: Heterogeneous network embedding representation for disease gene prediction // IEEE Journal of Biomedical and Health Informatics. 2019. V. 23. N 4. P. 1805–1815. https://doi.org/10.1109/JBHI.2018.2870728
19. Grover A., Leskovec J. node2vec: Scalable feature learning for networks // Proc. of the KDD'16 . International Conference on

References

1. Henaff M., Bruna J., LeCun Y. Deep convolutional networks on graph-structured data. *arXiv*, 2015, arXiv:1506.05163. https://doi.org/10.48550/arXiv.1506.05163
2. Wang X., Bo D., Shi C., Fan S., Ye Y., Yu P.S. A survey on heterogeneous graph embedding: methods, techniques, applications and sources. *IEEE Transactions on Big Data*, 2023, vol. 9, no. 2, pp. 415–436. https://doi.org/10.1109/TBDATA.2022.3177455
3. Shao B., Li X., Bian G. A survey of research hotspots and frontier trends of recommendation systems from the perspective of knowledge graph. *Expert Systems with Applications*, 2021, vol. 165, pp. 113764. https://doi.org/10.1016/j.eswa.2020.113764
4. László L. Random walks on graphs: a survey. *Combinatorics*. V. 2. 1993, pp. 1–46.
5. Li L., Wang Y., An L., Kong X., Huang T. A network-based method using a random walk with restart algorithm and screening tests to identify novel genes associated with Menière's disease. *PLOS ONE*, 2017, vol. 12, no. 8, pp. e0182592. https://doi.org/10.1371/journal.pone.0182592
6. Muslu Ö., Hoyt C.T., Lacerda M., Hofmann-Apitius M., Frohlich H. GuiltyTargets: Prioritization of novel therapeutic targets with network representation learning. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2022, vol. 19, no. 1, pp. 491–500. https://doi.org/10.1109/TCBB.2020.3003830
7. Li Y., Kuwahara H., Yang P., Song L., Gao X. PGCN: Disease gene prioritization by disease and gene embedding through graph convolutional neural networks. *biorxiv.org*, 2019. https://doi.org/10.1101/532226
8. Dutta A., Alcaraz J., TehraniJamsaz A., Cesar E., Sikora A., Jannesari A. Performance optimization using multimodal modeling and heterogeneous GNN. *arXiv*, 2023, arXiv.2304.12568. https://doi.org/10.48550/arXiv.2304.12568
9. Thanapalasingam T., van Berkel L., Bloem P., Groth P. Relational graph convolutional networks: Closer Look. *PeerJ Computer Science*, 2022, vol. 8, pp. e1073. https://doi.org/10.7717/PEERJ-CS.1073
10. Wang X., Ji H., Shi C., Wang B., Ye Y., Cui P., Yu P.S. Heterogeneous graph attention network. *Proc. of the WWW '19: The World Wide Web Conference*, 2019, pp. 2022–2032. https://doi.org/10.1145/3308558.3313562
11. Ali A., Bagchi A. An overview of protein-protein interaction. *Current Chemical Biology*, 2015, vol. 9, no. 1, pp. 53–65. https://doi.org/10.2174/221279680901151109161126
12. Malone J., Holloway E., Adamusiak T., Kapushesky M., Zheng J., Kolesnikov N., Zhukova A., Brazma A., Parkinson H. Modeling sample variables with an experimental factor ontology. *Bioinformatics*, 2010, vol. 26, no. 8, pp. 1112–1118. https://doi.org/10.1093/bioinformatics/btq099
13. Lee J., Yoon W., Kim S., Kim D., Kim S., So C.H., Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 2020, vol. 36, no. 4, pp. 1234–1240. https://doi.org/10.1093/bioinformatics/btz682
14. Zhang X., Zou Y., Shi W. Dilated convolution neural network with LeakyReLU for environmental sound classification. *Proc. of the 22nd International Conference on Digital Signal Processing (DSP)*, 2017. https://doi.org/10.1109/ICDSP.2017.8096153
15. Piñero J., Queralt-Rosinach N., Bravo A., Deu-Pons J., Bauer-Mehren A., Baron M., Sanz F., Furlong L.I. DisGeNET: A discovery platform for the dynamical exploration of human diseases and their genes. *Database*, 2015, vol. 2015. https://doi.org/10.1093/database/bav028
16. Alam M., Cevallos B., Flores O., Lunetto R., Yayoshi K., Woo J. Yelp Dataset Analysis using Scalable Big Data. *arXiv*, 2021, arXiv.2104.08396v1. https://doi.org/10.48550/arXiv.2104.08396
17. Li Y., Guo X., Lin W., Zhong M., Li Q., Liu Z., Zhong W., Zhu Z. Learning dynamic user interest sequence in knowledge graphs for click-through rate prediction. *IEEE Transactions on Knowledge and Data Engineering*, 2023, vol. 35, no. 1, pp. 647–657. https://doi.org/10.1109/TKDE.2021.3073717
18. Kuo Y., Wang R., Liu G., Shu Z., Wang N., Zhang R., Yu J., Chen J., Li X., Zhou X. HerGePred: Heterogeneous network embedding representation for disease gene prediction. *IEEE Journal of Biomedical and Health Informatics*, 2019, vol. 23, no. 4, pp. 1805–1815. https://doi.org/10.1109/JBHI.2018.2870728
19. Grover A., Leskovec J. node2vec: Scalable feature learning for networks. *Proc. of the KDD'16 . International Conference on*

Knowledge Discovery & Data Mining. 2016. P. 855–864. https://doi.org/10.1145/2939672.2939754
20. Yuxiao D., Chawla N., Swami A. metapath2vec: Scalable representation learning for heterogeneous networks // Proc. of the KDD'17 . International Conference on Knowledge Discovery & Data Mining. 2017. P 135–144. https://doi.org/10.1145/3097983.3098036
21. Mikolov T., Chen K., Corrado G., Dean J. Efficient estimation of word representations in vector space // Proc. of the Workshop ICLR. 2013.
22. Perozzi B., Al-Rfou R., Skiena S. DeepWalk: Online learning of social representations // Proc. of the KDD'14. 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2014. P. 701–710. https://doi.org/10.1145/2623330.2623732
23. Hu Z., Dong Y., Wang K., Sun Y. Heterogeneous graph transformer // Proc. of the WWW '20. The Web Conference. 2020. P. 2704–2710. https://doi.org/10.1145/3366423.3380027
24. He M., Huang C., Liu B., Wang Y., Li J. Factor graph-aggregated heterogeneous network embedding for disease-gene association prediction // BMC Bioinformatics. 2021. V. 22. P. 165. https://doi.org/10.1186/s12859-021-04099-3

Knowledge Discovery & Data Mining, 2016, pp. 855–864. https://doi.org/10.1145/2939672.2939754
20. Yuxiao D., Chawla N., Swami A. metapath2vec: Scalable representation learning for heterogeneous networks. *Proc. of the KDD'17 . International Conference on Knowledge Discovery & Data Mining*, 2017, pp 135–144. https://doi.org/10.1145/3097983.3098036
21. Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space. *Proc. of the Workshop ICLR*, 2013.
22. Perozzi B., Al-Rfou R., Skiena S. DeepWalk: Online learning of social representations. *Proc. of the KDD'14. 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014, pp. 701–710. https://doi.org/10.1145/2623330.2623732
23. Hu Z., Dong Y., Wang K., Sun Y. Heterogeneous graph transformer. *Proc. of the WWW '20. The Web Conference*, 2020, pp. 2704–2710. https://doi.org/10.1145/3366423.3380027
24. He M., Huang C., Liu B., Wang Y., Li J. Factor graph-aggregated heterogeneous network embedding for disease-gene association prediction. *BMC Bioinformatics*, 2021, vol. 22, pp. 165. https://doi.org/10.1186/s12859-021-04099-3

Авторы

Сидоренко Денис Александрович — аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, https://orcid.org/0009-0004-0571-5192, denisissveta@mail.ru
Шалыто Анатолий Абрамович — доктор технических наук, профессор, главный научный сотрудник, профессор, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, sc 56131789500, https://orcid.org/0000-0002-2723-2077, shalyto@mail.ifmo.ru

Authors

Denis A. Sidorenko — PhD Student, ITMO University, Saint Petersburg, 197101, Russian Federation, https://orcid.org/0009-0004-0571-5192, denisissveta@mail.ru
Anatoly A. Shalyto — D.Sc., Full Professor, Chief Scientific Researcher, ITMO University, Saint Petersburg, 197101, Russian Federation, sc 56131789500, https://orcid.org/0000-0002-2723-2077, shalyto@mail.ifmo.ru

Статья поступила в редакцию 23.04.2024
Одобрена после рецензирования 25.06.2024
Принята к печати 24.07.2024

Received 23.04.2024
Approved after reviewing 25.06.2024
Accepted 24.07.2024