Методы выявления новых тем научных публикаций на основе машинного обучения

Гавердовская Елизавета, 525 группа

7 ноября 2025 г.

Актуальность темы

Проблема: быстрый рост объема научных статей создает вызов для анализа и обработки новой информации, так как ежегодно публикуются около миллиона работ.

Задача: выявление новых перспективных, междисциплинарных и революционных направлений до того, как они получат широкое признание.

Постановка задачи

Дано:

1) множество научных публикаций $D = \{D_1, D_2, D_3, \dots, D_n\}$, n - количество опубликованных статей;

2) метаданные каждой статьи D_i: автор, название работы, год публикации, цитируемая литература, полный текст работы.

Цели задачи:

- поиск функции F, которая описывает модель машинного обучения и которая для момента времени t прогнозирует появление новой темы T_{new};
- поиск неожиданных комбинаций существующих тем , которые с наибольшей вероятностью сформируются ко времени t_1.

Представление данных

1. Узлы (вершины): авторы, названия статей, концепты, ключевые слова;
2. Ребра (связи): цитирование, соавторство, принадлежность к теме.

$G_t = (V, E_t)$ – неориентированный граф, где V – множество вершин, $E_t \subseteq V \times V$ – множество наблюдаемых связей до момента времени t.

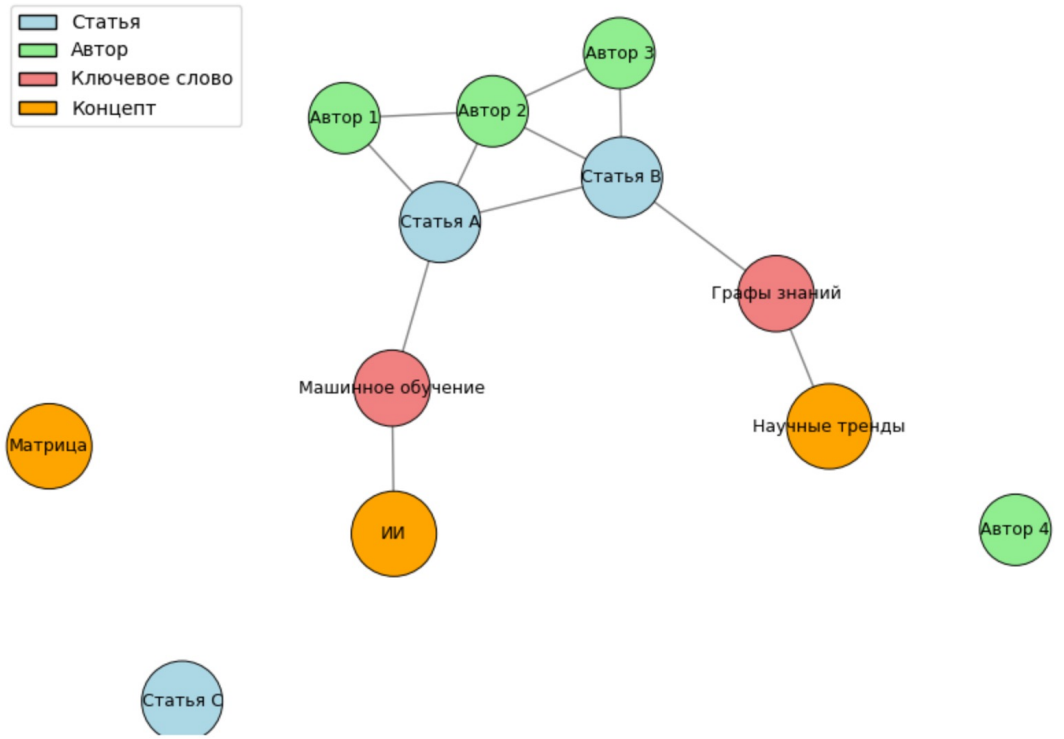


Рисунок 1 – Пример графа

Метрики оценки качества

Для прогноза новых тем для научных статей:
1) Precision (точность) – доля корректно найденных новых тем среди всех найденных:

$$Precision = \frac{TP}{TP + FP}.$$

2) Recall (полнота) – доля реально существующих новых тем, которые модель смогла обнаружить:

$$Recall = \frac{TP}{TP + FN}.$$

3) F1-score – используется при несбалансированных данных:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}.$$

4) ROC-AUC – способность модели отличать новые темы от старых по все порогам:

$$ROC - AUC = \int_0^1 TPR(FPR) d(FPR).$$

TP – число корректно выявленных тем; *FP* – число неверно выявленных тем; *FN* – число пропущенных новых тем; *TPR* - Recall; $FPR = \frac{FP}{FP+TN}$.

Извлечение признаков

Текстовые признаки

TF-IDF, word2vec, BERT – эмбеддинги текста, или способ преобразования объектов, такие как слова, в числовые векторы фиксированной размерности.

Тематические признаки

LDA/Topic Modeling – метод обучения без учителя, предназначенный для обнаружения скрытых абстрактных тем в больших коллекциях текстовых документов.

Графовые признаки

1. centrality – важность узла;
2. pagerank – ранжирование документов/мера центральности;
3. community detection – группировка узлов в кластере.

Алгоритмы для извлечения графовых признаков

1) Centrality определяет меру важности узла в графе при суммирование всех его входящих и исходящих связей.

$$C_i = \sum_j w_{ij}\text{, где } w_{ij} - \text{вес ребра между узлами } i \text{ и } j.$$

2) PageRank – это алгоритм ранжирования, который измеряет важность и авторитетность статьи на основе анализа ее ссылочного профиля. Учитывается количество ссылок и их качество.

$$PR(a) = (1 - d) + d\sum_{i=1}^{n}\frac{PR(T_i)}{C(T_i)}\text{, где}$$

$PR(a)$ – pagerank документа a; d – коэффициент затухания, который имитирует вероятность перехода к статье (~ 0.85); n –общее количество страниц, ссылающихся на работу a; T_i - каждая из страниц, ссылающихся на работу a; $PR(T_i)$ – pagerank одной страницы; $C(T_i)$ – общее количество источников на страницы T_i, которые ссылаются на другие источники.

3) Community detection – метод анализа сетей, который заключается группировке узлов в кластеры на основе их внутренних связей. Одна из метрик качества разбиения узлов – модулярность Q.

$$Q = \sum_{i,j}\left[\frac{A_{ij}}{2m} - \frac{k_i k_j}{4m^2}\right]\sigma(c_i, c_j)\text{, где}$$

m – количество связей; A – матрица смежности графа; k_i - степень вершины i; c_i - номер класса, к которому принадлежит вершина I; $\sigma(c_i, c_j) = \begin{cases} 1, c_i = c_j \\ 0, c_i \neq c_j \end{cases}$.

Основные подходы для прогнозирования

1. Алгоритм для прогнозирования «неожиданных комбинаций»
2. Машинное обучение на развивающихся графах знаний, например, темпоральные графовые нейронные сети
3. Ансамблевый подход, который комбинирует в себе несколько методов: например, нейронная сеть и ансамбль деревьев решений

Основные подходы для прогнозирования. Представление входных и выходных данных

1. Анализ неожиданных комбинаций.

- Входные данные: содержание (S), контекст (C)
- Выходные данные: мера неожиданности U

2. Темпоральные графовые нейронные сети.

- Входные данные: узлы графа (V), ребра ($E_t \subseteq V \times V$)
- Выходные данные: вероятность p_i, что в момент времени t появится новая тема

3. Ансамблевый подход.

- Входные данные: для ветки GNN (графовые признаки), для GBDT ветки (текстовые признаки)
- Выходные данные: предсказание (0 или 1), что в публикациях возникнет новая тема

Анализ неожиданных комбинаций

Данный метод основан на том, что научные прорывы часто возникают не результате углубления в одну область, а при комбинации знаний из далеких друг от друга дисциплин. Кроме того, он позволяет находить скрытые связи в большом объеме информации.

Компоненты комбинации:

1. содержание (content): концепт, темы статьи и др. (*S*);
2. контекст (context): место, в котором опубликована данная статья (*C*).

Задача метода: выявить и количественно оценить «неожиданность» комбинации (*S, C*) и связать ее с будущим влиянием работы.

В этом исследовании используются: 19 916 562 биомедицинских статей, опубликованных в период с 1865 по 2009 год из базы данных MEDLINE; 541 448 статей, опубликованных в период с 1893 по 2013 год в области физических наук из журналов, опубликованных Американским физическим обществом (APS), и 6 488 262 патентов, выданных в период с 1979 по 2017 год из базы данных патентов США

Shi F., Evans J. A. – 2023
GitHub: https://github.com/KnowledgeLab/hyper-novelty .

Анализ неожиданных комбинаций

Мера неожиданности: если тема S редко встречается в контексте C, их комбинация будет неожиданна.

1. Для дискретных событий (для конкретной темы):

KL-дивергенция (расстояние Кульбака-Лейблера) – это метрика для сравнения двух распределений вероятностей, показывающая, насколько одно распределение отличается от другого.

$$U(S,C) = \sum_i A(i) \log \frac{A(i)}{B(i)},$$

A(i) – истинная вероятность события *i*; *B(i)* – базовое распределение.

2. Для множества тем:

$$U(S,C) = 1 - \cos\left(V_s, V_{global}\right), \text{где}$$

V_s – вектор распределения всех тем в контексте *C*;

V_{global} – вектор глобального распределения всех тем;

Косинусное расстояние между векторами распределения:

$$\cos(A,B) = \frac{A \cdot B}{\|A\|\|B\|}$$

Интерпретация: чем выше *U*, тем необычнее сочетание темы и контекста.

Результаты прогнозирования

Новые контекстные комбинации также предсказуемы:

1. Биомедицина: AUC = 0,99

интерпретация: практически идеальное предсказание, область сильно зависит от заимствования методов извне (информатика, химия)

2. Физика: AUC = 0,88

интерпретация: высокая точность, скорее всего, область хорошо совместима с компьютерными технологиями

3. Изобретения: AUC = 0,83.

интерпретация: хорошая точность

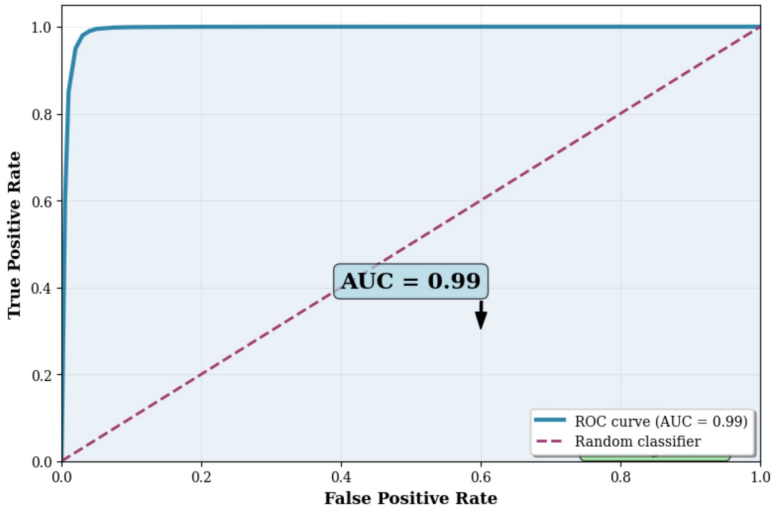


Рисунок 2 – График ROC-AUC биомедицины

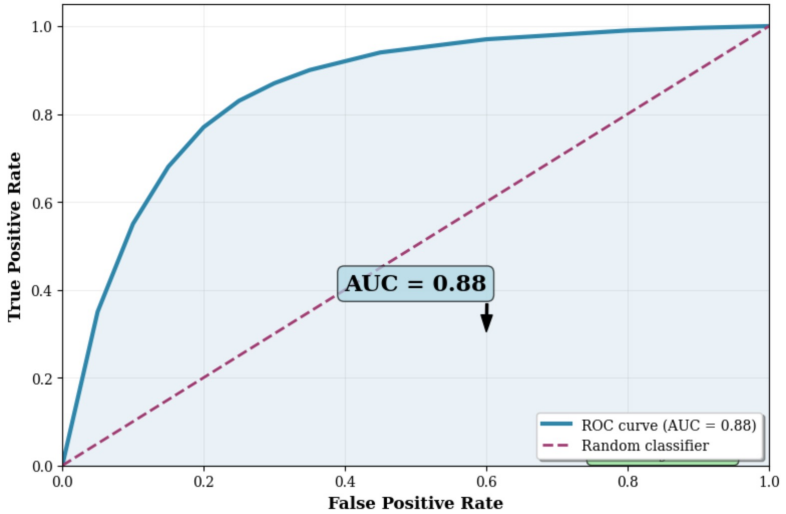


Рисунок 3 – График ROC-AUC физики

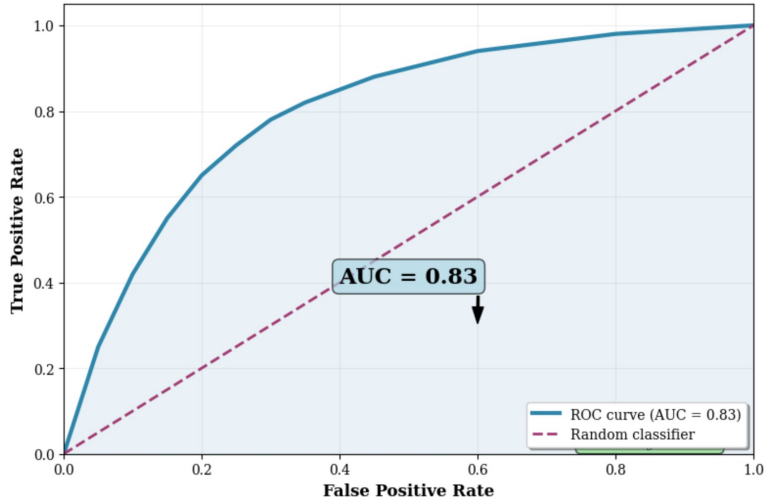


Рисунок 4 – График ROC-AUC изобретений

Графовые нейронные сети, GNN

Графовые нейронные сети, GNN (Graph Neural Networks) – тип нейронных сетей, предназначенных для работы с данными, представленными в виде графов, где узлы – объекты, а ребра – связи между ними. Работают на основе принципов передачи сообщений (message passing): агрегация информации от соседей и обновление собственного состояния узла.

$$h_v^{(k)} = Update\left(h_v^{(k-1)}, Aggregate\left(\left\{Message\left(h_u^{(k)}\right) \middle| u \in N(v)\right\}\right)\right), \text{где}$$

Message(…) – генерация сообщений от всех соседей; *Aggregate(…)* – агрегация этих сообщений в один вектор; *update(…)* – обновление собственного состояния узла $h_v^{(k-1)}$ с использованием агрегированной информации для получения нового состояния $h_v^{(k)}$.

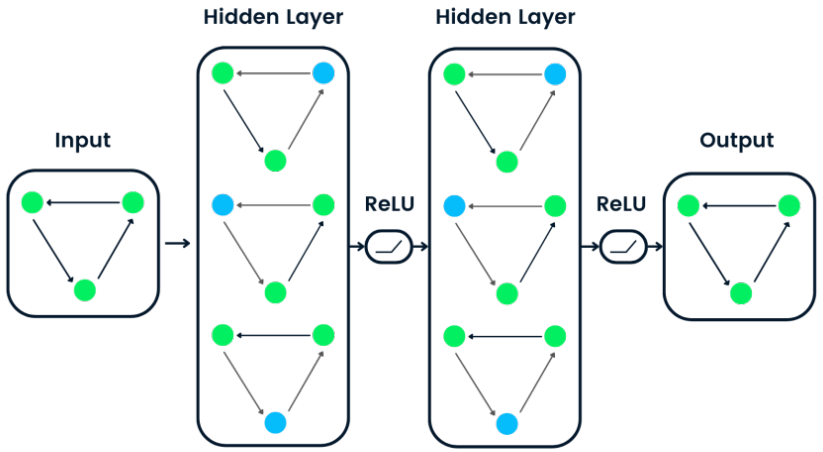


Рисунок 5 – Пример GNN

Gu X., Krenn M. – 2025
GitHub: https://github.com/artificial-scientist-lab/Impact4Cast

Подвид GNN - TGGN

Темпоральные графовые нейронные сети – это модели, которые комбинируют возможности графовых нейронных сетей и рекуррентных сетей, обрабатывающих временные последовательности.

Алгоритм работы этого метода:

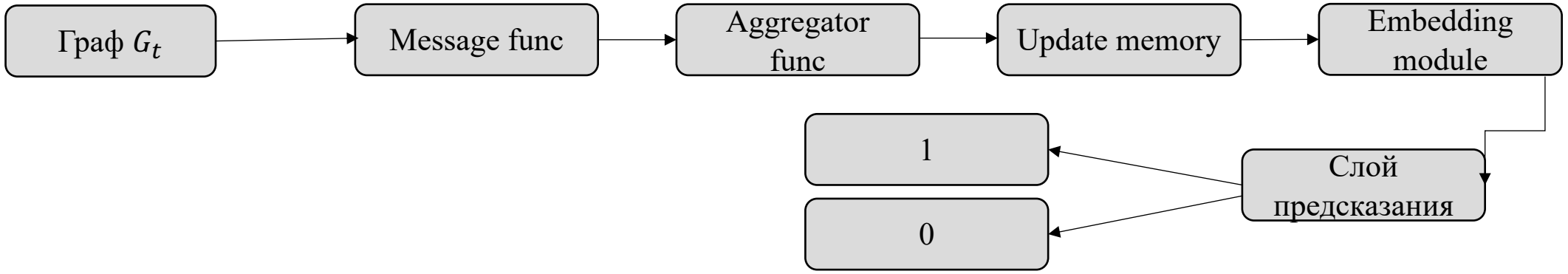
1. графовая часть: вначале GNN обрабатывает структуру графа, агрегируя информацию от соседних узлов, чтобы создать их представления;
2. временная часть: полученные представления на первом этапе обрабатываются по временной последовательности с помощью рекуррентных блоков*, которые позволяют модели запоминать и учитывать предыдущее состояние графа**.

*например, LSTM – используется для долгосрочной зависимости, или GRU – похож на LSTM, но имеет более легкую структуру.

**состояние графа – это эмбеддинг, то есть векторное представление узла, которое генерируется определенными функциями.

Обучение TGNN

Архитектура TGNN:



На шаге **Message func→Aggregator func** передается эмбеддинг узла (векторное представление, отражающее его связи на момент t); на шаг **Update memory** передается темпорально-обогащенный эмбеддинг (вектор, учитывающий историю графа до t; на шаг **Embedding module** передается финальный эмбеддинг для предсказания; на шаге **«слой предсказания»** прогнозируется вероятность или класс ответа и передается в вывод.

В слое предсказания TGNN используется функция активации и функция потерь для бинарной классификации – Binary Cross-Entropy (BCE). Она определяет разницу между прогнозируемым и фактическим значением.

$$BCE = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log p_i + (1 - y_i)\log(1 - p_i)], \text{где}$$

N – общее количество данных; y_i- это значение каждого элемента в наборе данных (0 или 1); p_i - это вероятность того, что предсказанное значение принадлежит к классу 1.

Результаты прогнозирования

Датасет включает в себя более 21 миллиона научных работ, с 1709 года (начиная с письма Антони ван Леувенхука) и по апрель 2023 года. Но в данной работе модель была обучена на 10 млн случайных данных из датасета с 2016 года по 2019.

ROC-AUC = 0.747 на тестовой выборке показывает, что модель обладает умеренной точностью, так как результат значительно выше случайного угадывания.
Также можно заметить, что на валидационной выборке метрика хуже, чем на тестовой, это может указывать на переобучение.

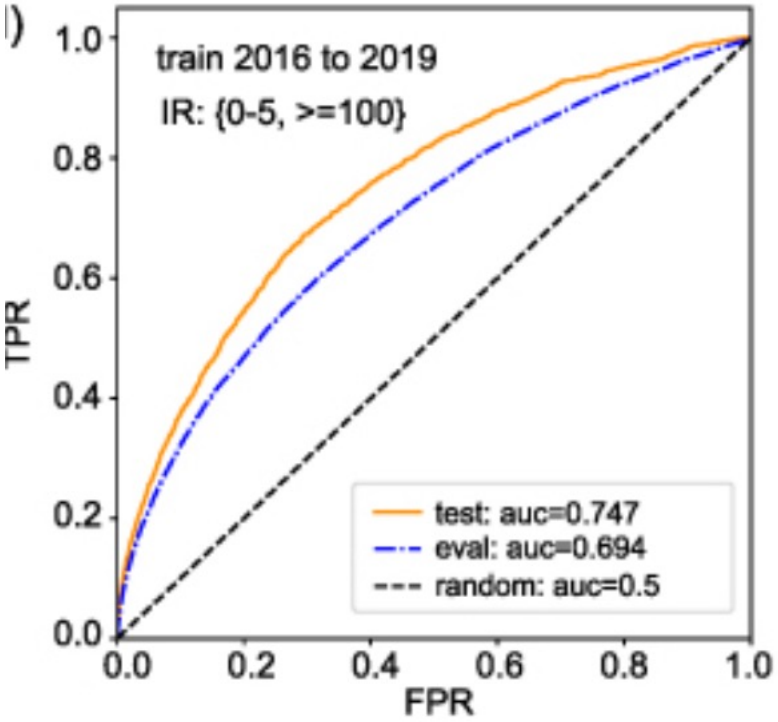


Рисунок 6 – ROC-AUC

Гибридные и ансамблевые подходы

Идея метода: использование комбинаций моделей, работающих с разными типами признаков, для увеличения точности предсказания.

Ключевые аспекты:

1) Комбинация моделей – объединение разных алгоритмов для улучшения предсказаний.

2) Работа с разными типами признаков – текстовые, числовые, сетевые и др.

3) Повышение точности и устойчивости – уменьшение ошибок и переобучения.

4) Популярные техники:

- Bagging (Random Forest)
- Boosting (XGBoost, LightGBM)
- Stacking (нейросети + классические модели)

GBDT – gradient boosting decision trees

GBDT – это метод машинного обучения, которое последовательно строит деревья, каждое из которых исправляет ошибки предыдущего.

$$F_M(x) = F_0(x) + \sum_{m=1}^{M} \gamma_m h_m(x)\,, \text{где}$$

$F_M(x)$ - итоговая модель после M итераций; $F_0(x)$ - начальное предсказание; γ_m - шаг обучения (learning rate) на итерации m; $h_m(x)$ – модель, обученная на итерации m.

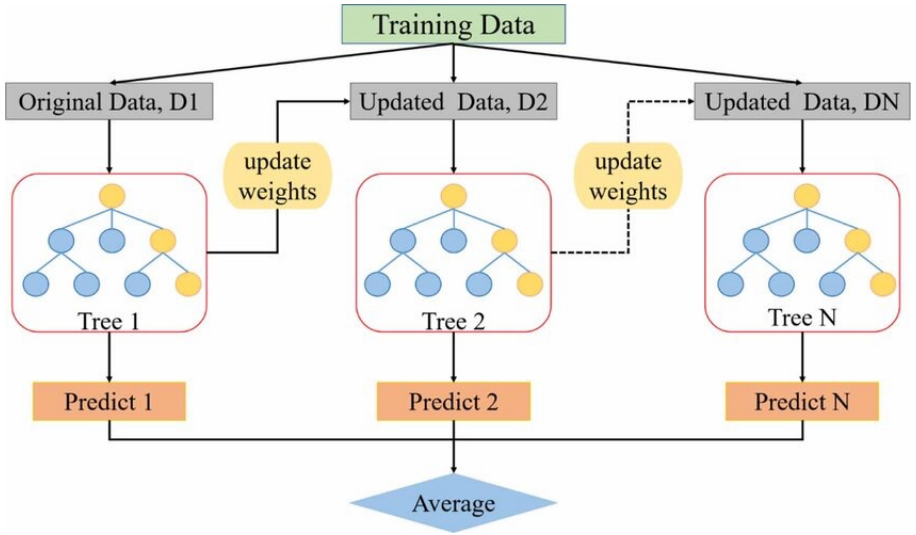


Рисунок 7 – Пример GBDT

Lu Y. Predicting Research Trends in Artificial Intelligence with Gradient Boosting Decision Trees and Time-aware Graph Neural Networks. - 2022

Архитектура гибридных моделей

1) GNN-ветвь:

- выявляет **временные паттерны**: как темы развиваются со временем;
- генерирует эмбеддинги для каждой публикации, отражающие её контекст в графе;
- улавливает **структурные зависимости;**

2) GBDT-ветвь:

- использует **атрибутивные признаки публикаций**: ключевые слова, количество авторов, журналы, год публикации;
- хорошо выявляет **корреляции и важность признаков** для классификации или прогнозирования новых тем.

3) Слияние моделей:

- **конкатенация эмбеддингов**: объединение выходов GNN и GBDT перед финальной классификацией;
- **стэкинг**: усреднение или взвешивание предсказаний двух моделей для итогового решения;
- позволяет использовать сильные стороны обеих ветвей одновременно.

Сравнительный анализ

	Метод неожиданных комбинаций	Temporal Graph Neural Networks	Гибридный подход
Основная идея	Выявляет новые темы через редкие или необычные сочетания ключевых слов/концепций в публикациях	Моделирует граф цитирований с учетом времени для прогнозирования новых тем	Комбинирует графовую обработку и анализ атрибутов публикаций для предсказания новых тем
Преимущества	Простота в реализации, хорошо выявляет радикально новые идеи	Учитывает структурные и временные зависимости	Использует сильные стороны обеих ветвей, повышает точность, так как учитывает и структуру и признаки публикаций
Ограничения	Игнорирует графовые и временные зависимости, чувствителен у шуму в текстах	Высокие вычислительные затраты, сложно масштабировать	Сложность реализации, требует огромных ресурсов для обучения и слияния двух моделей

Текущие ограничения

1) Масштабируемость: анализ длинных последовательностей или непрерывных потоков событий означает обработку и хранение очень больших объемов информации.

2) Во втором методе TGNN ключевым ограничением является потеря долгосрочной информации, может снижать точность выявления трендов.

3) Адаптация к новым трендам: модели плохо справляются с быстро меняющимися научными направления без переобучения.

4) Интерпретируемость: гибридные модели сложно объяснить, что затрудняет понимание причин выделения новых тем.

Заключение

Для выявления новых научных тем статей были использованы такие **методы**, как:

1. Алгоритм «неожиданных комбинаций»
2. Темпоральные графовые нейронные сети
3. Ансамблевый подход

Результатами исследования являются:

- Биомедицина: ROC-AUC = 0,99; физика: ROC-AUC = 0,88; изобретения: ROC-AUC = 0,83.
- ROC-AUC = 0.747 на тестовой выборке (размер тестовой выборки – 10 млн данных)

Перспектива использования данных направлений заключается в том, что комбинация двух подходов демонстрирует хороший результат прогнозирования в узкоспециализированных областях и сохраняет робастность при масштабе до 10 млн данных, что создает основу для системы прогнозирования научных трендов.

Литература

1. Shi F., Evans J. A. Surprising combinations of research contents and contexts are related to impact and emerge with scientific outsiders from distant disciplines // Nature Communications. — 2023.
2. Gu X., Krenn M. Forecasting high-impact research topics via machine learning on evolving knowledge graphs.
3. Lu Y. Predicting Research Trends in Artificial Intelligence with Gradient Boosting Decision Trees and Time-aware Graph Neural Networks.
4. Xiaomei Bai, Hui Liu, An Overview on Evaluating and Predicting Scholarly Article Impact - 2017
5. Томилова Н.И, Базаров С.А. Развитие систем прогнозирования на основе нейронных сетей - 2024