Прогнозирование количественного выхода химических реакций с помощью графовых нейронных сетей

Гунаев Руслан Гуламович

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

Консультант Ф. Никитин

Курс: Численные методы обучения по прецедентам (практика, В. В. Стрижов)/Группа 774, весна 2020

Прогнозирование количественного выхода химической реакции

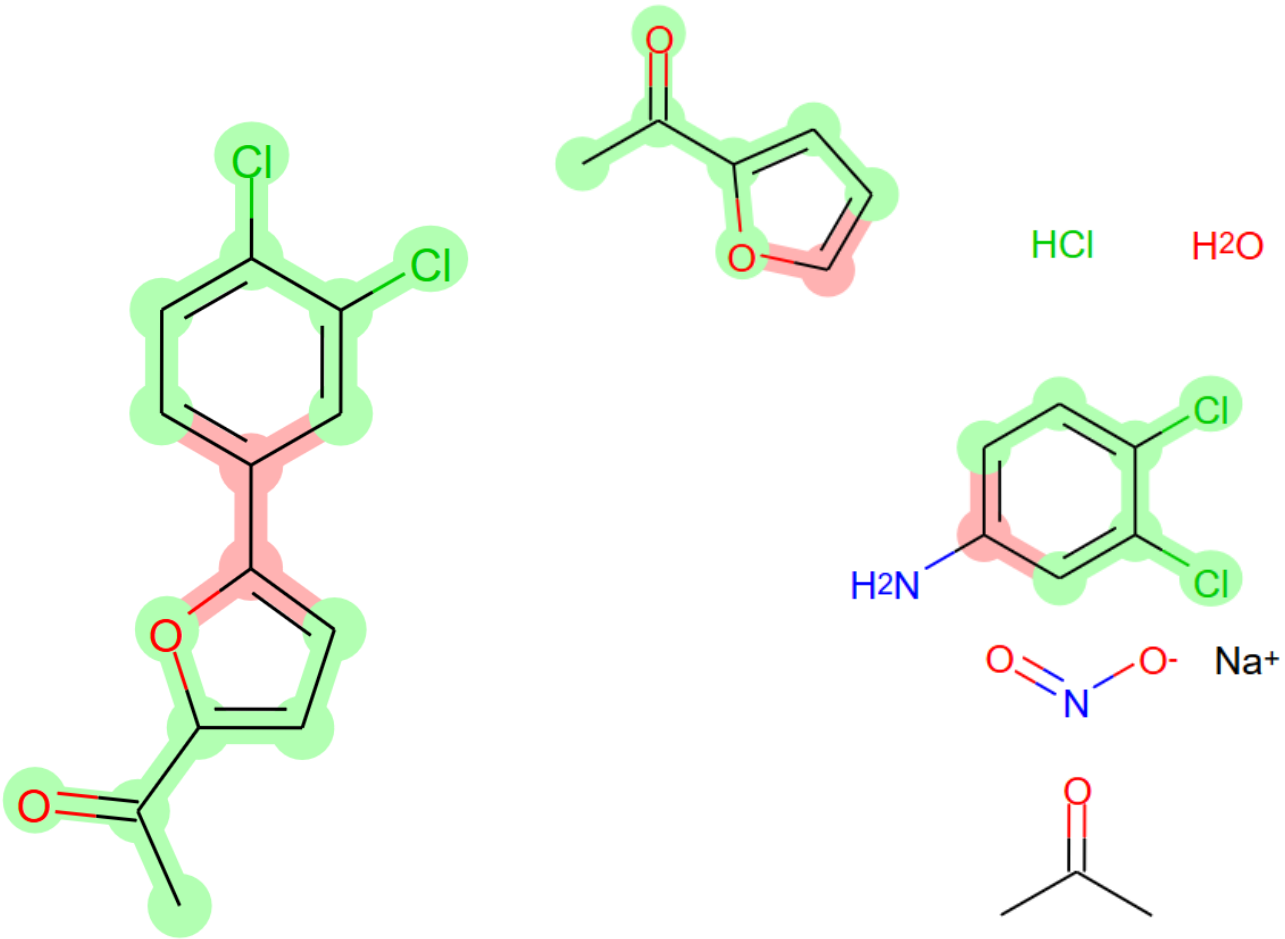
Цель

Предложить графовые нейронные сети для решения задачи регрессии на множестве молекулярных графов для прогнозирования количественного выхода химической реакции, используя экспертные знания.

Определение

Молекулярный граф – связный неориентированный граф, находящийся во взаимно-однозначном соответствии со структурной формулой химического соединения таким образом, что вершинам графа соответствуют атомы молекулы, а рёбрам графа — химические связи между этими атомами.
Экспертные знания – информация о химичических связях(одинарная, двойная, тройная, ароматическая) и о свойствах атомов(степень, явная валентность, гибридизация, неявная валентность, ароматичность, неявность и т.д.).

Продукт и реагенты химической реакции



(a) Продукт

(b) Реагенты

Описание данных

База реакций

1. 1 млн. реакций в формате SMARTS
2. Разделены продукты и реагенты
3. Известен основной продукт
4. Для 20% реакций известен количественный выход основного продукта

Справка

SMILES – язык, позволяющий однозначно закодировать молекурлярный граф строкой символов ASCII, SMARTS – поднастройка SMILES.
Реагент – вещество, участвующее в химической реакции.
Продукт – вещество, которое поменяло свое строение в результате реакции.
Основной продукт – молекула, включающая в себя наибольшее количество атомов среди всех продуктов реакции.

Литература

Junying Li, Deng Cai, and Xiaofei He.
Learning graph-level representation for drug discovery.
arXiv preprint arXiv:1709.03741, 2017.

Michael Schlichtkrull, Thomas N Kipf, et al.
Modeling relational data with graph convolutional networks.

In *European Semantic Web Conference*, pages 593–607.
Springer, 2018.

Постановка задачи

Выборка

$X = \{g_i, y_i\}_{i=1}^N$, $g_i \in G$ – множество входных молекулярных графов, y_i – выход реакции.

Модель

$$f(g, \mathbf{W}) : \; G \times \Omega \to \mathbb{R},$$

Ω – пространство параметров модели.

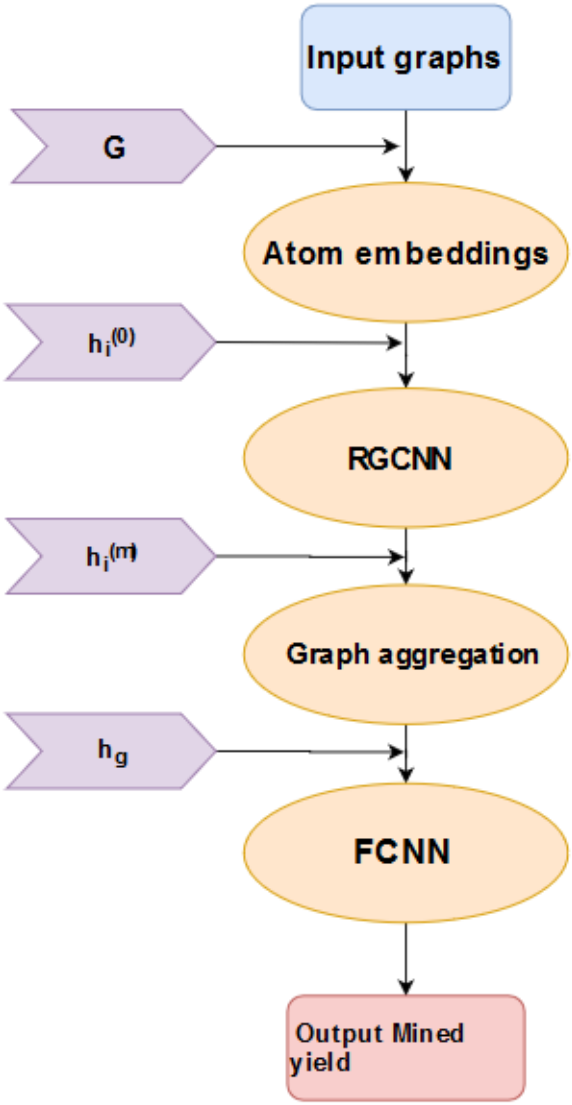
Задача

Найти $\mathbf{W}^*$ такую, что

$$\mathbf{W}^* = \arg\min_{\mathbf{W} \in \Omega} \frac{1}{N} \sum_{i=1}^{N} \|y_i - f(g_i, \mathbf{W})\|_2^2.$$

Архитектура модели

1. Инициализация векторных представлений вершин молекулярного графа
2. Графовая сверточная нейронная сеть
3. Агрегация графа
4. Полносвязная нейронная сеть
5. Получение выхода реакции



Инициализация вершин молекулярного графа

Эмбеддиниги

$$\mathbf{h}_{ik}^{(0)} = W_i^k,$$

W^k – матрица эмбеддинга для k-го категориального признака, i – номер столбца матрицы, верхний индекс $\mathbf{h}_{ik}^{(0)}$ означает, что вектор на нулевом слое.

Представление атома

$$\mathbf{h}_i^{(0)} = \text{concat}[\mathbf{h}_{i1}^{(0)}, h_{i2}^{(0)}, \ldots, \mathbf{h}_{iK}^{(0)}],$$

K – количество категориальных признаков.

Графовая сверточная нейронная сеть

RGCNN

$$\mathbf{h}_i^{(l+1)} = \mathrm{ReLU}\left(\mathbf{W}^{(l)}\mathbf{h}_i^{(l)} + \sum_{r\in R}\sum_{j\in N_i}\frac{1}{c_{i,r}}\mathbf{W}_r^{(l)}\mathbf{h}_j^{(l)}\right),$$

R – множество типов ребер графа(типов химических связей), $\mathbf{W}, \mathbf{W}_r$ – параметры модели, $\mathbf{h}_i^{(l)}$ – векторное представление a_i атома на l слое, $c_{i,r}$ – нормировочный коэффициент.

Модель расширенного графа

Обновление векторных представлений вершин

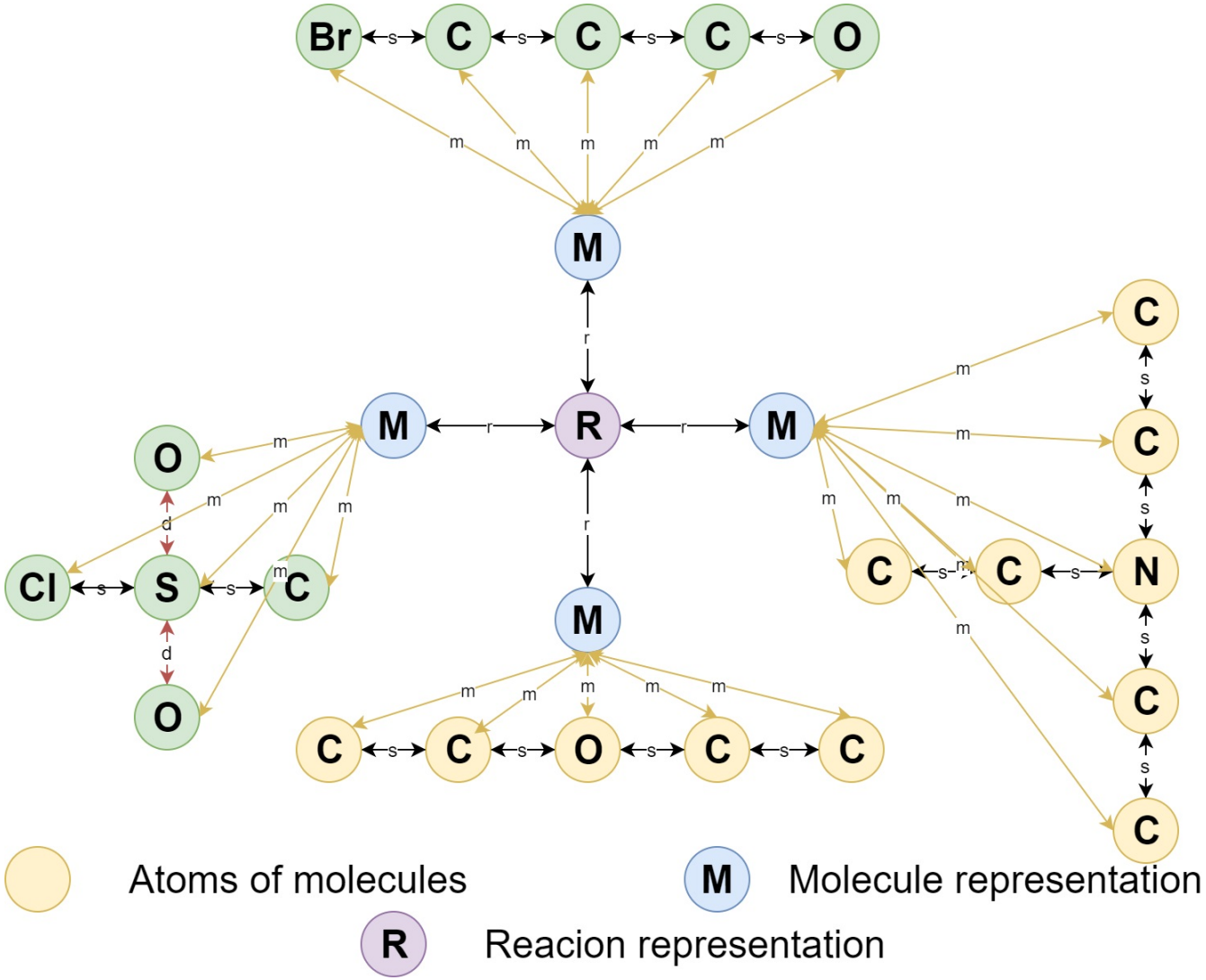
$$\mathbf{h}_i^{(l+1)} = \mathrm{ReLU}\left(\mathbf{W}^{(l)}\mathbf{h}_i^{(l)} + \mathbf{W}_{ml}^{(l)}\mathbf{h}_{m_k}^{(l)} + \sum_{r\in R}\sum_{j\in N_i}\frac{1}{c_{i,r}}\mathbf{W}_r^{(l)}\mathbf{h}_j^{(l)}\right),$$

$$\mathbf{h}_{m_k}^{(l+1)} = \mathrm{ReLU}\left(\mathbf{W}^{(l)}\mathbf{h}_{m_k}^{(l)} + \mathbf{W}_{rl}^{(l)}\mathbf{h}_r^{(l)} + \sum_{j\in m_k}\frac{1}{|m_k|}\mathbf{W}_{ml}^{(l)}\mathbf{h}_j^{(l)}\right),$$

$$\mathbf{h}_r^{(l+1)} = \mathrm{ReLU}\left(\mathbf{W}^{(l)}\mathbf{h}_r^{(l)} + \sum_{m_j\in M}\frac{1}{|M|}\mathbf{W}_{rl}^{(l)}\mathbf{h}_{m_j}^{(l)}\right),$$

$\mathbf{W}_{rl}$ – матрица преобразований типа реакция-молекула,
$\mathbf{W}_{ml}$ – молекула-атом,
$\mathbf{h}_i^{(l)}$ – векторное представление атома,
$\mathbf{h}_r^{(l)}$– векторное представление реакции,
$\mathbf{h}_{m_k}^{(l)}$ – векторное представление молекулы.

Расширенный молекулярный граф



Расширенный граф с введенными вершинами — представлениями молекул и реакции.

Полносвязная сеть

Агрегация графа

$$\mathbf{h}_g = \frac{1}{n}\sum_{i=1}^{n}\mathbf{h}_{i=1}^{(m)},$$

$\mathbf{h}_g$ – векторное представление расширенного графа, m – число слоев RGCNN.

FCNN

$$\mathbf{h}_g^{(l+1)} = \mathrm{ReLU}(\mathrm{linear}(\mathbf{h}_g^{(l)})),$$

$\hat{y} = \mathrm{linear}(\mathbf{h}_g^{(t)})$, $\hat{y}$ – выход сети.

Функция ошибки

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\|y - \hat{y}\|_2^2,\ y \text{ – реальный выход реакции.}$$

Вычислительный эксперимент

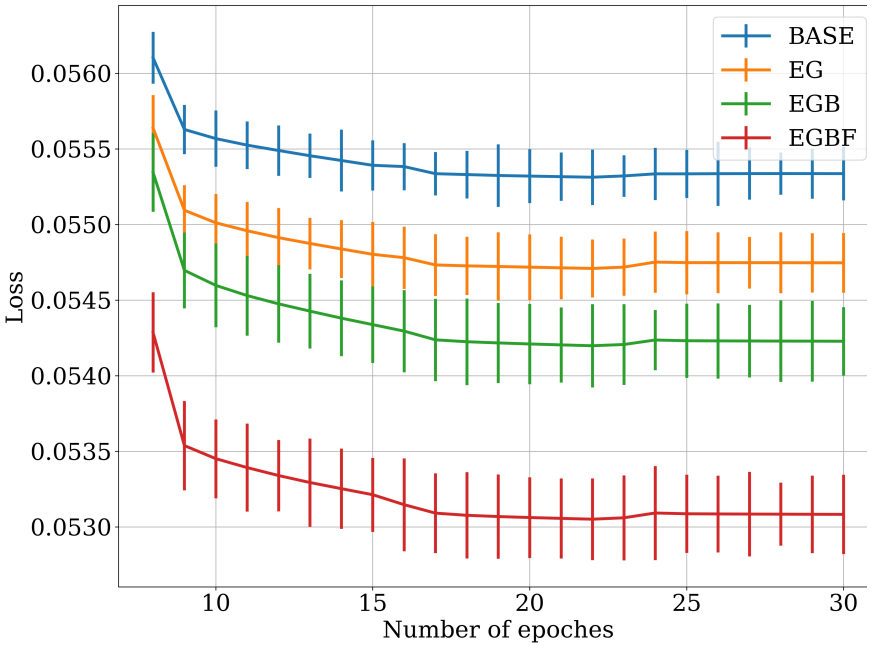
Цели эксперимента

1. Проверить, повышается ли качество модели при последовательном добавлении дополнительной информации о структуре графа.
2. Может ли предлагаемая модель давать более высокое качество, чем константная модель.

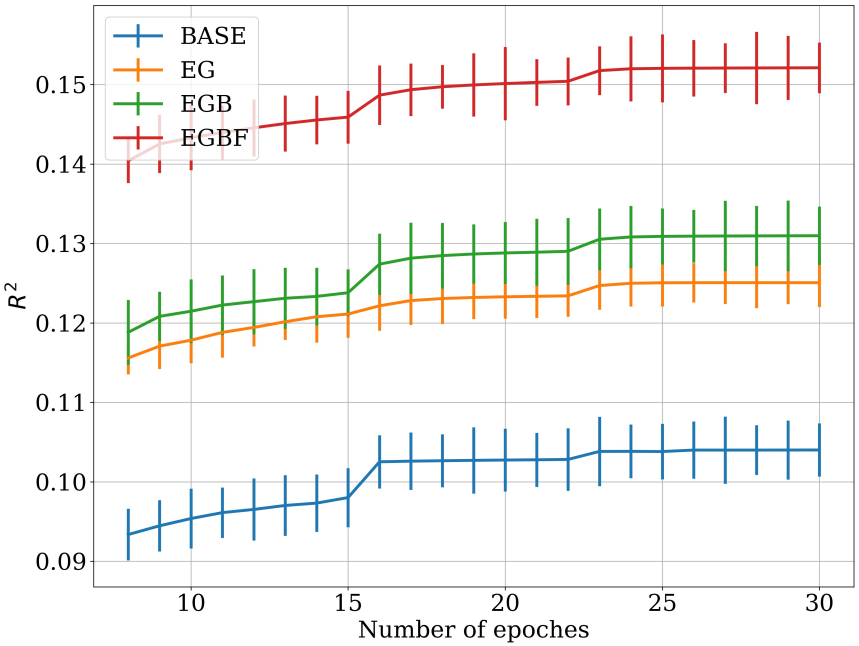
Эксперименты

1. **CONST** – константная модель,
2. **BASE** – базовая модель(RGCNN + FCNN),
3. **EG** – модель расширенного графа,
4. **EGB** – модель с различными типами химической связи,
5. **EGBF** – модель с дополнительной информацией о свойтсвах атомов.

Результаты экспериментов



Ошибка во время обучения



R^2 для тестовой выборки

При каждой модификации архитектуры уменьшаются ошибки во время обучения и повышается качество на тестовой выборке.

Вывод

Модель	**Mined Yield**	
	R^2	MAE
CONST	0	0.211 ± 0.004
BASE	0.104 ± 0.002	0.198 ± 0.003
EG	0.125 ± 0.003	0.194 ± 0.002
EGB	0.131 ± 0.006	0.186 ± 0.002
EGBF	0.152 ± 0.005	0.174 ± 0.003

R^2 – коэффициент детерминации. MAE – среднее абсолютное отклонение между реальными выходами реакций и предсказанными на тестовой выборке.

Каждая из предложенных моделей показала более высокое качество чем константная модель. Самое высокое качество наблюдается у модели **EGBF**.

Заключение

Полученные результаты

1. При добавлении дополнительной информации о структуре графа качество модели повысилось.
2. Предложенная модель показала более высокое качество, чем константная модель.

Дальнейшие исследования

1. Классификация типов реакций для повышения качества регрессии.
2. Определение влияния свойств атомов на качество предложенной модели.