Использование графовых моделей для прогнозирования продуктов химических реакций

Гунаев Руслан [1]***, Никитин Филипп***[1]***, Стрижов Вадим***[1]
gunaev.rg@phystech.edu
Московский физико-технический институт

Решается задача прямого синтеза химических элементов. Требуется определить количественный выход основного продукта химической реакции по молекулам исходных веществ. Данная задача является одной из ключевых для автоматизации процессов разработки лекарств, а именно входит в более общую проблему ретросинтеза химических элементов. Предложенная модель использует экспертные знания для решения задачи. Качество решения задачи прямого синтеза измеряется по полному совпадению предсказанного и оригинального молекулярных графов основного продукта.

Ключевые слова: *ретросинтез маленьких молекул, прямой синтез, молекулярный граф.*

1 Введение

Исследование в области химинформатики можно разделить на две части: получение молекул - кандидатов, получение способов синтеза данного кондидата из известных, получаемых ранее веществ. Суть первой задачи заключается в создании таких молекулярных структур, например в виде молекулярного графов, которые, предположительно, обладают полезными свойствами: они не токсичны, взаимодействуют с другими соединениями. Имеется множество исследований, направленых на автоматизацию данного процесса. В работе [12] были обозрены существовавшие концепты, нацеленные на использование вычислительных ресурсов для ускорения открытия новых веществ. Развитие машинного обучения повлияло на развитие вычислительной химии и биологии [2]. В частности, предприняты попытки генерировать предположения с помощью рекурентныйх нейронных сетей [8, 10], которые посимвольно генерировали строковое представление молекулы SMILES [15]. Также для этой задачи применяются вариационные автокодировщики [5] и графовые нейронные сети [6, 7].

Цель второй задачи — это создание способа получения предлагаемой молекулы из уже ранее полученных веществ путем реализации цепочки реакций. При этом число реакций в цепочке не должно быть велико. Первые попытки автоматизировать решение задачи синтеза и ретросинтеза веществ были предприняты в 1969 [4].

Сейчас активно развиваются методы глубинного обучения для решения поставленной задачи: предложен вариант использования neural machine translation [1]: продемонстрированы результаты модели перевода с рекурентными слоями [13], а также Transformer модели [14]. В работе [3] была построена графовая сверточная сеть, которая оценивала вероятность образования связи для заданных вершин.

В данной работе рассматриваться задача прямого синтеза. А именно рассматривается задача, в которой требуется по описаниям молекул исходных веществ определить молекулу основного продукта химической реакции как результата химического взаимодействия исходных веществ. В качестве представления молекул используются молекулярные графы. Поставленная задача решается как задача бинарной классификации вершин(атомов) в графе исходных веществ в два этапа. На пермом этапе атомы классифицируются по признаку принадлежности основному продукту. На втором этапе выделяются атомы, которые

в процессе реакции изменили "локальную конфигурацию". Под локальной конфигурацией подразумевается совокупность вершины и смежных с ней ребер (типов связи).

В работе предложен метод классификации вершин в молеклярных графах исходных веществ — графе, состоящем из нескольких компонент. Решение базируется на модели Relational Graph convolution neural network [11]. Предложено несколько модификаций, позволяющих использовать модель для классификации вершин в несвязанном графе. Более того данные модификации позволяют эффективно моделировать межмолекулярное взаимодействие, играющее ключевую роль в протекании реакций. Модель RGNN является обобщеним модели graph convolution neural network [9] для графов с различными типами ребер. В нашем случае это очень важно, так как ребрами в молекулярном графе являются типы химической связи: одинарная, двойная, тройная, ароматическая и тд.

2 Постановка задачи

В работе решается задача регрессии для количественного определения основного продукта химической реакции. Исходные вещества и реагенты могут быть представлены в виде нескольких молекулярных графов. В результате взаимодействия этих веществ образуются новые вещества. Представленный набор молекулярных графов исходных веществ и реагентов представляет собой множество неориентированных графов с различными типами ребер и вершин. Тип вершины – номер химического элемента, тип ребра – тип химической связи.

2.1 Расширенное представление графа

Для каждой молекулы добавим новую вершину и соединим ее с каждым атомом этой молекулы ребром особого типа. Далее создадим еще одну вершину и соединим ее со всеми ранее добавленными вершинами ребрами того же типа.

2.2 Начальное скрытое представление атома

Каждый атом представлен несколькими категориальными признаками, такими как валентность, тип атома и др. Признаки будем обозначать $k_n, n \in \{1, \dots K\}$.

Для каждого из представленных категориальных признаков строится векторное вредставление (embedding). В формуле (1) используются обознаения: i — номер вершины в графе, m_i — тип вершины, k — номер признака, W_k — матрица весов для k-го категориального признака.

$$\mathbf{h}_i^{(0)k} = W_{m_i}^k \tag{1}$$

Итоговое преставление атома есть конкатенация эмбедингов $\mathbf{h}_i^{(0)k}$ по всем признакам (2).

$$\mathbf{h}_i^{(0)} = [\mathbf{h}_i^{(0)0}, \mathbf{h}_i^{(0)1}, \mathbf{h}_i^{(0)2}, \dots, \mathbf{h}_i^{(0)K}] \tag{2}$$

2.3 Обновление скрытых состояний атома

В графовой сверточной нейронной сети обновление скрытых состояний происходит в соответствии с формулой (3). В этой формуле: σ — нелинейность, N_i — множество индексов атомов смежных с i-м, c_i — нормализационный фактор.

$$\mathbf{h}_i^{(l+1)} = \sigma\left(W^{(l)}\mathbf{h}_j^{(l)} + \sum_{j \in N_i} \frac{1}{c_i} W^{(l)}\mathbf{h}_j^{(l)}\right) \tag{3}$$

Недостатком данной модели является предположение о том, что все связи между вершинами эдентичны. Для предсказаний высокой точности необходимо, чтобы модель выучивала типы связей в процессе обучения.В связи с тем, что информация о типе связи между атомами известна, то подход выглядит недостаточно рациональным. Эту проблему решает RGCNN, обновление скрытых состояний в которой происходит в соответствии с формулой (4). В данной модели смежные с заданной вершиной разным типом связи атомы учитываются с разными весами.

$$\mathbf{h}_{(l+1)}^{i} = \sigma\left(W^{(l)}\mathbf{h}^{(l)} + \sum_{r\in R}\sum_{j\in N_i}\frac{1}{c_{i,r}}W_r^{(l)}\mathbf{h}_j^{(l)}\right) \tag{4}$$

Литература

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[2] Keith T Butler, Daniel W Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh. Machine learning for molecular and materials science. *Nature*, 559(7715):547, 2018.

[3] Connor W Coley, Wengong Jin, Luke Rogers, Timothy F Jamison, WH Green, R Barzilay, KF Jensen, et al. A graph-convolutional neural network model for the prediction of chemical reactivity. 2018.

[4] EJ Corey and W Todd Wipke. Computer-assisted design of complex organic syntheses. *Science*, 166(3902):178–192, 1969.

[5] Hanjun Dai, Yingtao Tian, Bo Dai, Steven Skiena, and Le Song. Syntax-directed variational autoencoder for structured data. *arXiv preprint arXiv:1802.08786*, 2018.

[6] Nicola De Cao and Thomas Kipf. Molgan: An implicit generative model for small molecular graphs. *arXiv preprint arXiv:1805.11973*, 2018.

[7] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. In *Advances in neural information processing systems*, pages 2224–2232, 2015.

[8] Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276, 2018.

[9] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[10] Marcus Olivecrona, Thomas Blaschke, Ola Engkvist, and Hongming Chen. Molecular de-novo design through deep reinforcement learning. *Journal of cheminformatics*, 9(1):48, 2017.

[11] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *European Semantic Web Conference*, pages 593–607. Springer, 2018.

[12] Gisbert Schneider and Uli Fechner. Computer-based de novo design of drug-like molecules. *Nature Reviews Drug Discovery*, 4(8):649, 2005.

[13] Philippe Schwaller, Theophile Gaudin, David Lanyi, Costas Bekas, and Teodoro Laino. "found in translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chemical science*, 9(28):6091–6098, 2018.

[14] Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Costas Bekas, and Alpha A Lee. Molecular transformer for chemical reaction prediction and uncertainty estimation. *arXiv preprint arXiv:1811.02633*, 2018.

[15] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

Поступила в редакцию