Предсказание свойств и типов атомов в молекулярных графах при помощи сверточных сетей*

Манучарян В., Грудинин С., Кадукова М., Стрижов В., В
manucharyan.va@phystech.edu
Московский физико-технический институт (государственный университет), Москва

Статья посвящена определению типов атомов в молекулярных графах. В существующих на данный момент моделях тип определяется с помощью цепочки условных операторов на основе предсказанной гибридизации. Для автоматизации этой процедуры в статье предлагается учитывать 3D структру молекулярного графа и использовать свёрточную нейронную сеть на молекулярных графах. Для вычислительного эксперимента используется база данных PDBBindCN, в котором определены тип атома (около 150 классов), гибридизация атома (4 класса) и тип связи (5 классов). Результат работы нового алгоритма сравнивается с Knodle [1].

Ключевые слова: *машинное обучение, классификация, свёрточные нейронные сети, молекулярные графы.*

1 Введение

1.1

В этой работе решается задача классификации типов атомов. Атомы разделены на типы, исходя из их химических свойств и их окружения, например, sp3-гибридизованный углерод. Типы атомов применяются в предсказании взаимодействия молекул и в вычислительных методах в медицине и биологии, например, в виртуальном скрининге при разработке новых лекарств [2].

Молекула представлена в виде трёхмерного молекулярного графа — связного неориентированного графа, в котором вершины — атомы, рёбра — химические связи. Обычно он представляется в виде планарного графа, в этой работе граф трёхмерный. Такое представление учитывает трёхгранные углы, образованные атомами, и расположение атомов оносительно друг друга в пространстве. Таким образом для каждого атома учитывается информация о соседних атомах, от которой зависит тип атома.

Предложены как модели, использующие простые геометрические соображения [3], функциональные группы [4], гибридизацию и заряд атома [5–8], так и модели, основанные на свёрточных нейронных сетях (cNN) на графах [15, 16, 20]. Подробнее о том, как введена опреация конволюции в данных работах в секции "Обзор существующих операций конволюции".

В данной работе предлагается cNN, в архитектуре которой определены как слои, характеризующие признаки атомов, так и слои, характеризующие признаки пар атомов. Благодаря такой архитектуре, операция конволюции учитывает атом вместе с его окружением. Это позволяет определять тип атомов напрямую без длинной цепочки условных операторов на основе гибридизации атома.

Предложенный алгоритм сравнивается с алгоритмом, реализованным в библиотеке по распознаванию типов атомов Knodle [1], основанным на мультиклассовой классификации при помощи метода опорных векторов.

*Научный руководитель: Стрижов В. В. Задачу поставил: Стрижов В. В. Консультанты: Грудинин С., Кадукова М.

1.2 Обзор существующих операций конволюции

1. Задан молекулярный граф $G(V, E)$. В нём выбирается произвольный набор вершин V', где $|V'| = n$, а n — параметр сети. Далее для каждой вершины v из V' строится граф-фильтр размера k. Изначально этот граф пустой и строится по следующему принципу. Добавим в него v. Затем будем добавлять вершины на расстоянии 1 от v (но не более k), если их меньше k, то добавляем вершины на расстоянии 2 от v, и так далее, пока не наберём k вершин или пока нечего будет добавлять. Далее нормализуем каждый граф-фильтр, т.е. нумеруем вершины в некотором порядке и строим вектор на основе этой нумерации, и обучаем нейронную сеть на полученных векторах. [15] Но данный метод подходит скорее для определения свойств молекулы в целом, а не отдельных атомов.

2. Заданы граф $G(V, E)$, и каждая веришна $v_i \in V$ ассоциирована с вектором признаков $X_i \in \mathbb{R}^m$, или $\mathbf{X} = (X_1^T, \ldots, X_n^T) \in \mathbb{R}^{n \times m}$. Определим матрицу смежности $\mathbf{A}$ и матрицу существования пути длины k $\widetilde{\mathbf{A}}^{(k)}$. Тогда $(\mathbf{A}^k)_{i,j}$ — количество путей длины k между v_i и v_j, а

$$\widetilde{\mathbf{A}}^{(k)} = min\{\mathbf{A}^k + I, 1\},$$

где минимум берётся поэлементно. Определим матрицу параметров сети $\mathbf{W}^{(k)}$, определим адаптивный фильтр

$$\widetilde{\mathbf{W}}^{(k)} = \mathbf{G} \circ \mathbf{W}^{(k)},$$

где $\circ$ определяет поэлементное матричное уможение, а

$$\mathbf{G} = \text{sigmoid}([\widetilde{\mathbf{A}}^{(k)}, \mathbf{X}] \cdot \mathbf{Q}),$$

где $[\cdot, \cdot]$ — конкатенация матриц, $\mathbf{Q} \in M_{n+m,n}$ — матрица параметров, делающая размерность $\mathbf{G}$такой, как у $\mathbf{A}$. Таким образом фильтр будет учитывать как признаки вершины, так и её окружение. Определим операцию конволюции:

$$L^{(k)} = (\widetilde{\mathbf{W}}^{(k)} \circ \widetilde{\mathbf{A}}^{(k)})\mathbf{X} + B_k,$$

где B_k - вектор смещения. Получая на вход матрицу признаков $\mathbf{X} \in \mathbb{R}^{n \times m}$, сеть выдаёт матрицу признаков $L = [L^{(1)}, \ldots, L^{(K)}] \in \mathbb{R}^{n \times mK}$, таким образом, каждая вершина будет ассоциирована с вектором признаков, которые можно использовать для предсказания типов атомов. [20]

2 Постановка задачи

2.1 Описание выборки

Выборка содержит 15000 молекул в формате mol2 из базы данных PDBBindCN. Для каждой молекулы из базы построена матрица смежности и матрица длин кратчайших путей между атомами. Также с целью признакового описания атомов построена матрица длин связей, углов и двугранных углов. Для каждого атома определены следующие дескрипторы: название элемента, электроотрицательность, включение в кольцо, смешанное произведение векторов связи этой вершины. На основе этих данных молекула описана $N \times D$ матрицей, где каждая строчка соответствует атому, D — количество признаков, и $N \times N$ матрицами, для кажого парного признака задана своя матрица.

2.2 Постановка задачи определения типов атомов

Заданы $\mathfrak{G} = \{\mathfrak{s}_1, ..., \mathfrak{s}_m\}$ — множество типизированных атомов, $\mathbf{y} = [y_1, ..., y_m]$ — типы атомов. Задан $G = \{g_1, ..., g_D\}$ — набор функций, g_j отображает $\mathfrak{s}_i$ в (i, j) элемент матрицы

X:

$$g_j : (b_j, \mathfrak{s}_i) \mapsto x_{ij} \in \mathbb{R}^1,$$

где b_j — набор параметров g_j.

Определена модель f, сопостовляющая каждой строке $\mathbf{X}$ число из отрезка $[0,\ 1]$:

$$f(\mathbf{w}, \mathbf{X}) = \frac{1}{1 + \exp(-\mathbf{X}\mathbf{w})},$$

где оптимальные параметры $\hat{\mathbf{w}}$ минимизируют функцию потерь

$$\hat{\mathbf{w}} = arg \min_{\mathbf{w} \in \mathbb{R}^D} S(\mathbf{w}|f, \mathbf{X}, y),$$

где

$$S(\mathbf{w}|f, \mathbf{X}, y) = -\ln \left(\sum_{i=1}^{m} y_i \log f(x_i, \mathbf{w}) + (1 - y_i) \log\bigl(1 - f(x_i, \mathbf{w})\bigr) \right).$$

2.3 Архитектура сети

Параметрами сети являются: параметры функции f: $w_1, \ldots, w_n$, глубина сети, глубина конволюции, способ получения молекулярных признаков, метод оптимизации параметров

Сеть состоит из двух видов слоёв: атомного и парного. Первый слой — двумерная матрица, где каждому атому соответствует вектор признаков. Второй слой — трёхмерная матрица, где каждой паре атомов соответствует вектор признаков.

Пусть x — атомный слой, $\mathfrak{s}$ — атом, тогда обозначим $A^x_{\mathfrak{s}}$ — вектор признаков атома $\mathfrak{s}$ в слое x. Аналогично, y — парный слой, $(\mathfrak{s}_1, \mathfrak{s}_2)$ — пара атомов, тогда $P^y_{(\mathfrak{s}_1, \mathfrak{s}_2)}$ — вектор признаков пары $(\mathfrak{s}_1, \mathfrak{s}_1)$ в слое y.

Пусть $f(z) = zI(z > 0)$, $g(z_1, ..., z_n) = \sum_{i=1}^{n} z_i$, $x, x_1, ..., x_n$ - слои. Опишем несколько операций, с помощью которых можно получать новые слои:

– Новый атомный слой y из нескольких предыдущих атомных слоёв $x_i, i \in \overline{1, n}$:

$$A^y = (A \to A)(A^{x_1}, \ldots, A^{x_n}).$$

Для каждого атома $\mathfrak{s}$ определим

$$A^y_{\mathfrak{s}} = f(+ \sum_{i=1}^{n} w_i A^{x_i}_{\mathfrak{s}}),$$

– Новый парный слой y из нескольких предыдущих парных слоёв $x_i, i \in \overline{1, n}$:

$$P^y = (P \to P)(P^{x_1}, \ldots, P^{x_n}).$$

Для каждой пары вершин $(\mathfrak{s}_1, \mathfrak{s}_1)$ определим

$$P^y_{(\mathfrak{s}_1, \mathfrak{s}_1)} = f(+ \sum_{i=1}^{n} w_i P^{x_i}_{(\mathfrak{s}_1, \mathfrak{s}_1)}),$$

– Новый атомный слой y из предыдущего парного слоя x:

$$A^y = (P \to A)(P^x).$$

Для каждого атома $\mathfrak{s}$ определим

$$A^y_{\mathfrak{s}} = g(f(c + w_1 P^x_{(\mathfrak{s}, \mathfrak{s}_1)}), f(c + w_1 P^x_{(\mathfrak{s}, \mathfrak{s}_2)}), f(c + w_1 P^x_{(\mathfrak{s}, \mathfrak{s}_3)}), ...),$$

где g вычисляется для всех пар, содержащих $\mathfrak{s}$,

– Новый парный слой y из предыдущего атомного слоя x:

$$P^y = (A \to P)(A_x).$$

Для каждой пары вершин $(\mathfrak{s}_1, \mathfrak{s}_1)$ определим

$$P^y_{(\mathfrak{s}_1,\mathfrak{s}_1)} = g\left(f(c + w_1 A^x_{\mathfrak{s}_1} + w_2 A^x_{\mathfrak{s}_2}), f(c + w_1 A^x_{\mathfrak{s}_2} + w_2 A^x_{\mathfrak{s}_1})\right).$$

Опишем, как с помощью этих операций получить из k-ого атомного и парного слоёв $(k+1)$-ые. А именно, пусть заданы A^k — k-ый атомный слой и P^k — k-ый парный слой. Сначала построим вспомогательные слои $A^{k'}$, $A^{k''}$, $P^{k'}$, $P^{k''}$, необходимые только для конструирования $(k+1)$-ых слоёв:

$$A^{k'} = (A \to A)(A^k),\ A^{k''} = (P \to A)(P^k),\ P^{k'} = (A \to P)(A^k),\ P^{k''} = (P \to P)(P^k),$$

и используя их получаем A^{k+1}, P^{k+1}:

$$A^{k+1} = (A \to A)(A^{k'}, A^{k''}),\ P^{k+1} = (P \to P)(P^{k'}, P^{k''}).$$

Проделав эту процедуру несколько раз, получаем финальный атомный слой A. Далее следует полносвязный слой и softmax.

3 Вычислительный эксперимент

Для оценки ошибки предложенной модели и её сравнения с Knodle [1] проведён эксперимент. Свёрточная нейронная сеть реализована с помощью PyTorch, библиотеки с открытым исходным кодом для машинного обучения. Выборка разделена на обучающую и тестовую в соотношении 4:1. Сеть обучалась 2000 итераций, используя Adagrad с коэффициентом скорости обучения 0.003. Точность предсказания на тестовой выборке предложенного алгоритма — 45%, точность Knodle — 94%

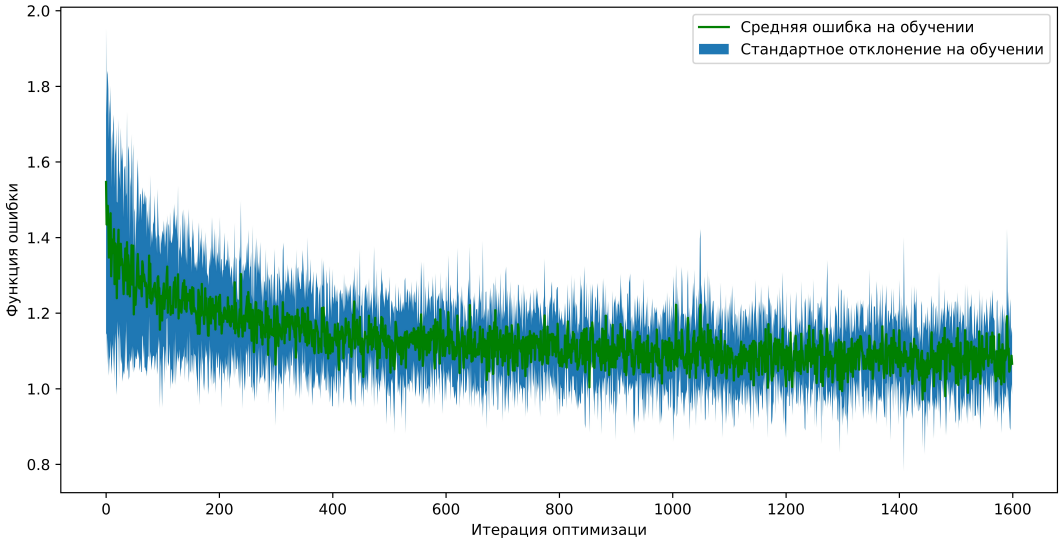

Рис. 1 Зависимость ошибки от итерации

Изменение функции ошибки на итерациях оптимизации показано на рисунке 2. Видим, что ошибка на обучающей выборке стабилизируется, стандртное отклонение невелико, что говорит о разумности нашей модели.

4 Заключение

В работе предложена модель, позволяющая определять тип атома автоматическив. С целью сравнения качества предоженной модели и модели, предсказывающей тип на основе гибридизации и цепочки условных операторов, проведён вычислительный эксперимент, демонстрирующий низкое качество классификация. Архитектуру сети следует пересмотреть и улучшить. Jupyter Notebook с реализацией сети находится в свободном доступе [21].

Литература

[1] *Maria Kadukova, Sergei Grudinin* Knodle: A Support Vector Machines-Based Automatic Perception of Organic Molecules from 3D Coordinates // Journal of Chemical Information and Modeling, American Chemical Society, 2016, 56 (8), pp.1410-1419.

[2] *Bohdan Waszkowycz, David E Clark, and Emanuela Gancia* Outstanding challenges in protein–ligand docking and structure-based virtual screening // Wiley Interdiscip. Rev.: Comput. Mol. Sci., 1(2):229-259, 2011.

[3] *Jon C Baber and Edward E Hodgkin* Automatic assignment of chemical connectivity to organic m the Cambridge structural database // J. Chem. Inf. Comput. Sci., 32(5):401–406, 1992.

[4] *Manfred Hendlich, Friedrich Rippmann, and Gerhard Barnickel* Bali: Automatic assignment of bond and atom types for protein ligands in the brookhaven protein databank // J. Chem. Inf. Comput. Sci., 37(4):774–778, 1997.

[5] *Elke Lang, Claus-Wilhelm von der Lieth, and Thomas Forster* Automatic assignment of bond orders based on the analysis of the internal coordinates of molecular structures // Anal. Chim. Acta, 265(2):283–289, 1992.

[6] *Yuan Zhao, Tiejun Cheng, and Renxiao Wang* Automatic perception of organic molecules based on essential structural information // J. Chem. Inf. Model., 47(4):1379–1385, 2007.

[7] *Daan MF van Aalten, R Bywater, John BC Findlay, Manfred Hendlich, Rob WW Hooft, and Gert Vriend* Prodrg, a program for generating molecular topologies and unique molecular descriptors from coordinates of small molecules // J. Comput.-Aided Mol. Des., 10(3):255–262, 1996.

[8] *Qian Zhang, Wei Zhang, Youyong Li, Junmei Wang, Liling Zhang, and Tingjun Hou* A rule based algorithm for automatic bond type perception // J. Cheminf., 4(1):1–10, 2012.

[9] *Paul Labute* On the perception of molecules from 3d atomic coordinates // J. Chem. Inf. Model., 45(2):215–221, 2005.

[10] *Gerd Neudertand Gerhard Klebe* Fconv: Format conversion, manipulation and feature computation of molecular data // Bioinformatics, 27(7):1021–1022, 2011.

[11] *Matheus Froeyen and Piet Herdewijn* Correct bond order assignment in a molecular framework using integer linear programming with application to molecules where only non-hydrogen atom coordinates are available // J. Chem. Inf. Model., 45(5):1267–1274, 2005.

[12] *Sascha Urbaczek, Adrian Kolodzik, Inken Groth, Stefan Heuser, and Matthias Rarey* Reading pdb: Perception of molecules from 3d atomic coordinates // J. Chem. Inf. Model., 53(1):76–87, 2012.

[13] *unmei Wang, Wei Wang, Peter A Kollman, and David A Case* Automatic atom type and bond type perception in molecular mechanical calculations // J. Mol. Graphics Modell., 25(2):247–260, 2006.

[14] *Anna Katharina Dehof, Alexander Rurainski, Quang Bao Anh Bui, Sebastian Bocker, Hans-Peter Lenhof, and Andreas Hildebrandt* Automated bond order assignment as an optimization problem // Bioinformatics, 27(5):619–625, 2011.

[15] *Mathias Niepert, Mohamed Ahmed, Konstantin Kutzkov* Learning Convolutional Neural Networks for Graphs // , 2016

[16] *Steven Kearnes, Kevin McCloskey, Marc Berndl, Vijay Pande, Patrick Riley* Molecular Graph Convolutions: Moving Beyond Fingerprints // , 2016

[17] *Duvenaud DK, Maclaurin D, Iparraguirre J, Bombarell R, Hirzel T, Aspuru-Guzik A, Adams RP* Convolutional networks on graphs for learning molecular fingerprints // Advances in neural information processing systems, pp 2224–2232, 2015

[18] *Lusci A, Pollastri G, Baldi P* Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules // J Chem Inf Model 53(7): 1563–1575, 2013

[19] *Merkwirth C, Lengauer T* Automatic generation of complementary descriptors with molecular graph network // J Chem Inf Model 45(5):1159–1168, 2005

[20] *Zhenpeng Zhou, Xiaocheng Li* Convolution on Graph: A High-Order and Adaptive Approach // , 2017

[21] *Манучарян В.* Реализация свёрточной нейронной сети для предсказания типов атомов // https://github.com/Intelligent-Systems-Phystech/Group594/blob/master/Manucharyan2018AtomicTypePredictionInUsingCNN/code/Manucharyan2018 2017

Поступила в редакцию