

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

Манучарян В., Грудинин С., Кадукова М., Стрижов В., В
manucharyan.va@phystec.edu

Московский физико-технический институт (государственный университет), Москва

Статья посвящена определению типов атомов и валентности в молекулярных графах. В существующих на данный момент моделях не учитывается 3D структура молекулярных графов, и тип определяется с помощью цепочки условных операторов на основе предсказанной гибридизации. Статья предлагает делать это при помощи свёрточной нейронной сети, обученной 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 эффективно классифицировать атомы, для предсказания взаимодействия молекул. При этом молекулы представлены в виде трёхмерных молекулярных графов, что применяется в вычислительных методах в медицине и биологии, виртуальный скрининг при разработке новых лекарств [2].

Молекулярный граф - это связный неориентированный граф, в котором вершины - атомы, рёбра - химические связи. Обычно он представляется в виде планарного графа, в этой же работе граф трёхмерный. Тип атома - это $\langle \dots \rangle$

На данный момент существуют как простые 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]. Так как тип для каждого атома зависит от его окружения, то свёртка с соседними атомами даст искомый результат.

В данной работе предлагается $\langle \dots \rangle$. Это позволяет определять тип атомов напрямую без длинной цепочки условных операторов на основе гибридизации атома.

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

1.2 Обзор операции конволюции

1. Задан молекулярный граф G , размер фильтра k и гиперпараметр n . В нём выбирается произвольный набор вершин V , где $|V| = n$. Далее для каждой вершины v из V строится "граф-фильтр" размера k . Этот граф 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

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

k вершин или пока нечего будет добавлять. Далее нормализуем каждый "граф-фильтр" и обучаем нейронную сеть на полученных векторах. [15] Но данный метод подходит скорее для определения свойств молекулы в целом, а не отдельных атомов.

2. Пусть заданы граф $G(V, E)$, и каждая вершина $v_i \in V$ ассоциирована с вектором признаков $X_i \in \mathbb{R}^m$, или $X = (X_1^T, \dots, X_n^T) \in \mathbb{R}^{n \times m}$. Определим матрицу смежности A . Тогда $A_{i,j}^k$ - количество путей длины k между v_i и v_j , а

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

где \min берётся поэлементно. Пусть W_k - матрица весов, определим адаптивный фильтр

$$\widetilde{W}_k = g \circ W_k,$$

где \circ определяет поэлементное матричное умножение, а

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

где $[\cdot, \cdot]$ - конкатенация матриц, $Q \in M_{n+m,n}$ - матрица параметров. Таким образом фильтр будет учитывать как признаки вершин, так и соседние вершины. Теперь всё готово для определения операции конволюции:

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

, где B_k - вектор смещения. Таким образом, получая на вход матрицу признаков $X \in \mathbb{R}^{n \times m}$, сеть выдаёт матрицу признаков $L = [L^{(1)}, \dots, L^{(K)}] \in \mathbb{R}^{n \times mK}$, т.е. 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 строим матрицу смежности G и матрицу расстояний D —длин кратчайших путей между атомами. Также строим матрицу 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 матрицей $N \times D$ (каждая строчка соответствует атому, D —количество признаков) и несколькими матрицами $N \times N$, для каждого парного признака задана своя матрица.

2.2 Формальная постановка задачи

Зададим $\mathfrak{G} = \{\mathfrak{s}_1, \dots, \mathfrak{s}_m\}$ —множество типизированных атомов, $\mathbf{y} = [y_1, \dots, y_m]$ —типы атомов.

Пусть $G = \{g_1, \dots, g_D\}$ —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 .

Определим модель f , сопоставляющую 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

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

2.3 Параметры сети

Параметрами сети являются:

- параметры функции f : w_1, \dots, w_n
- глубина сети
- глубина конволюции
- способ получения молекулярных признаков
- метод оптимизации параметров

2.4 Архитектура сети

В сети будет два вида слоёв: атомный и парный. Первый суть 2-мерная матрица, где каждому атому соответствует строка. Второй слой суть 3-мерная матрица, где каждой паре атомов соответствует строка.

Определение 2.1 Пусть x - атомный слой, \mathfrak{s} - атом, тогда $A_{\mathfrak{s}}^x$ - значение атома \mathfrak{s} в слое x . Аналогично, y - парный слой, $(\mathfrak{s}_1, \mathfrak{s}_2)$ - пара атомов, тогда $P_{(\mathfrak{s}_1, \mathfrak{s}_2)}^y$ — значение пары $(\mathfrak{s}_1, \mathfrak{s}_2)$ слое y .

Пусть $f(z) = zI(z > 0)$, $g(z_1, \dots, z_n) = \sum z_i$, x, x_1, \dots, x_n - слои. Опишем несколько операций, с помощью которых можно получать одни слои из других:

- $(A \rightarrow A): A_{\mathfrak{s}}^y = f(+ \sum_{i=1}^n w_i A_{\mathfrak{s}}^{x_i})$
- $(P \rightarrow P): P_{(\mathfrak{s}_1, \mathfrak{s}_2)}^y = f(+ \sum_{i=1}^n w_i P_{(\mathfrak{s}_1, \mathfrak{s}_2)}^{x_i})$
- $(P \rightarrow A): A_{\mathfrak{s}}^y = g(f(c + w_1 P_{(\mathfrak{s}, \mathfrak{s}_1)}^x), f(c + w_1 P_{(\mathfrak{s}, \mathfrak{s}_2)}^x), f(c + w_1 P_{(\mathfrak{s}, \mathfrak{s}_3)}^x), \dots)$, где g вычисляется для всех пар, содержащих \mathfrak{s}
- $(A \rightarrow P): P_{(\mathfrak{s}_1, \mathfrak{s}_2)}^y = g(f(c + w_1 A_{\mathfrak{s}_1}^x + w_2 A_{\mathfrak{s}_2}^x), f(c + w_1 A_{\mathfrak{s}_2}^x + w_2 A_{\mathfrak{s}_1}^x))$

Опишем, как с помощью этих операций получить из k -ого атомного и парного слоёв $(k+1)$ -ые. А именно, пусть есть A^k и P^k . Сначала получим промежуточные слои

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

и уже используя их получим $(k+1)$ -ые слои:

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

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

3 Вычислительный эксперимент

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 in 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 Hoof, and Gert Vriend* Prodr, 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 Neudert and 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] *Junmei 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

Поступила в редакцию