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 при помощи методов машинного обучения. Предлагается использовать свёрточные нейронные сети (cNN), обученные на 3D структуре молекулярных графов. Для обучения используется <...> датасет, в котором определены тип атома (около 150 классов), гибридизация атома (4 класса) и тип связи (5 классов). Результат работы нового алгоритма будет сравниваться с Knodle [1].

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

1 Введение

1.1

В этой работе мы хотим научиться эффективно классифицировать атомы, для того чтобы предсказывать взаимодействие молекул. Это важно во многих вычислительных методах в медицине и биологии, например, виртуальный скрининг при разработке новых лекарств [2]. При этом молекулы представимы в виде трёхмерных молекулярных графов, что позволяет использовать методы машинного обучения на графах.

“В последние годы были предложены несколько алгоритмов. Самые ранние работы основывались на простых геометрических соображениях, использующих длины связей и валентные углы [3]. Позже стали учитывать функциональные группы [4], гибридизацию и заряд атома [5–8]. Чтобы уменьшить влияние ошибок в экспериментальном определении структуры были использованы некоторые подходы: максимальное взвешенное паросочетание [9, 10], поиск структуры Льюиса [11], максимизация [12] или минимизация [13, 14] некой метрики.”

Также используют модели, основанные на свёрточных нейронных сетях (cNN) на графах [15, 16, 20]. Подробнее об этом в секции "Обзор операции конволюции".

В данной работе предлагается <..>.

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. Пусть заданы 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, b так далее, пока не наберём k вершин или пока нечего будет добавлять. Далее нормализуем каждый "граф-фильтр"и обучаем нейронную

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

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 веришна $v_i \in V$ ассоциирована с вектором признаков $X_i \in \mathbb{R}^m$, или $X = (X_1^T, \ldots, X_n^T) \in \mathbb{R}^{n \times m}$. 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 уможение, а

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

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 опреации 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 $L = [L^{(1)}, \ldots, L^{(K)}] \in \mathbb{R}^{n \times mK}$, т.е. каждая вершина будет ассоциирована с вектором признаков, которые можно использовать для предсказания типов атомов. [20]

2 Постановка задачи

2.1 Описание выборки

15000 молекул в формате mol2 из базы данных PDBBindCN. Для каждой молекулы определена матрица смежности G и матрица расстояний D (длины кратчайших путей между атомами). Так же имеются матрица длин связей, углов и двугранных углов. Для каждого атома определены следующие дескрипторы:

- название элемента
- электроотрицательность
- включение в кольце
- смешанное произведение векторов связи этой вершины

На основе этих данных молекулу можно описать матрицей NxD0 (каждая строчка соответствует атому, D0 - количество признаков) и несколькими матрицами NxN (для кажого парного признака своя матрица).

2.2 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, a - атом, тогда A_a^x - значение атома a в слое x. Аналогично, y - парный слой, (a, b) - пара атомов,тогда $P_{(a,b)}^y$ - значение пары (a, b) слое y.

Пусть $f(z) = zI(z > 0)$, $g(z_1, ..., z_n) = \sum z_i$, $x, x_1, ..., x_n$ - слои. Параметры: $c, w_1, \ldots, w_n$. Опишем несколько операций, с помощью которых можно получать одни слои из других:

- $(A \to A)$: $A_a^y = f(+\sum_{i=1}^n w_i A_a^{x_i})$
- $(P \to P)$: $P_{(a,b)}^y = f(+\sum_{i=1}^n w_i P_{(a,b)}^{x_i})$
- $(P \to A)$: $A_a^y = g(f(c+w_1 P_{(a,b)}^x), f(c+w_1 P_{(a,c)}^x), f(c+w_1 P_{(a,d)}^x), ...)$, где g вычисляется для всех пар, содержащих a
- $(A \to P)$: $P_{(a,b)}^y = g(f(c+w_1 A_a^x + w_2 A_b^x), f(c+w_1 A_b^x + w_2 A_a^x)))$

Лемма 1 Эти операции поддерживают следующий инвариант: если применить ко входу перестановку σ, то $\forall$ слоёв x, y: A^x, P^y переставляются согласно перестановке σ

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

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

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

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

Проделав эту процедуру несколько раз, получим финальный атомный слой A.

2.3 Параметры сети

Параметрами сети являются:

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

2.4 Формальная постановка задачи

Пусть $\mathfrak{G} = \{\mathfrak{s}_1, ..., \mathfrak{s}_\mathfrak{m}\}$ - множество атомов в различных молекулах. Пусть $\mathbf{y} = \{y_1, ..., y_m\}$ - типы атомов.

Пусть $G = \{g_1, ..., g_n\}$ - набор функций, таких что $\forall i \forall j$ g_j отображает $\mathfrak{s}_\mathfrak{i}$ в (i, j) элемент матрицы X:

$$g_j : (b_j, \mathfrak{s}_\mathfrak{i}) \to x_{ij} \in \mathbb{R}^1,$$

где b_j - набор параметров для g_j.

Определим модель f, сопостовляющую каждой строке X число из [0, 1]

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

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

$$\hat{w} = arg\min_w S(w|f, X, y),$$

где

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

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

Поступила в редакцию