

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

Манучарян Вардан

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 (сNN), обученные на 3D структуре молекулярных графов. Для обучения используется $\langle \dots \rangle$ 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 (сNN) на графах [15, 16].

В данной работе предлагается $\langle \dots \rangle$.

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 функций, используемых для конволюции

f - произвольная функция, например:

- $f(z) = \frac{1}{1+e^{-z}}$
- $f(z) = zI(z > 0)$
- $f(z) = \log(1 + \exp(x))$

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

- для $a < 0$ (подбирается в ходе обучения сети)

$$f(z) = \begin{cases} z & z > 0 \\ az & z < 0 \end{cases}$$

g - произвольная коммутативная функция, например:

- $g(z_1, \dots, z_n) = \sum z_i$
- $g(z_1, \dots, z_n) = \prod z_i$
- $g(z_1, \dots, z_n) = \max z_i$
- $g(z_1, \dots, z_n) = \frac{\sum z_i}{n}$

1.3 Обзор представления молекул

- Z-matrix
- матрица типов связи, матрица минимальных расстояний, матрица состояния в одном кольце вершин
-

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
- гибридизация
- тип
- тип2
- включение в кольцо
- смешанное произведение векторов связи этой вершины

На основе этих данных молекулу можно описать матрицей $N \times D_0$ (каждая строка соответствует атому, D_0 - количество признаков) и несколькими матрицами $N \times N$ (для каждого 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 $f(z) = zI(z > 0)$, $g(z_1, \dots, z_n) = \sum z_i$, x, x_1, \dots, x_n - слои. Параметры: c, w_1, \dots, w_n . Опишем несколько операций, с помощью которых можно получать одни слои из других:

- $(A \rightarrow A): A_a^y = f(+ \sum_{i=1}^n w_i A_a^{x_i})$
- $(P \rightarrow P): P_{(a,b)}^y = f(+ \sum_{i=1}^n w_i P_{(a,b)}^{x_i})$

- $(P \rightarrow A): A_a^y = g(f(c + w_1 P_{(a,b)}^x), f(c + w_1 P_{(a,c)}^x), f(c + w_1 P_{(a,d)}^x), \dots)$, где g вычисляется для всех пар, содержащих a
- $(A \rightarrow P): P_{(a,b)}^y = g(f(c + w_1 A_a^x + w_2 A_b^x), f(c + w_1 A_b^x + w_2 A_a^x))$

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

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

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

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

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

2.3 Параметры сети

Параметрами сети являются:

- параметры функции f : c, w_1, \dots, w_n
- глубина сети
- глубина конволюции (?непонятно что это)
- способ получения молекулярных признаков
- метод оптимизации параметров

2.4 Формальная постановка задачи

Пусть $\mathfrak{S} = \{\mathfrak{s}_1, \dots, \mathfrak{s}_m\}$ - множество атомов в различных молекулах. Пусть $\mathbf{y} = \{y_1, \dots, y_l\}$ - типы атомов.

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

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

где b_j - набор параметров для g_j .

Определим

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

где оптимальные параметры \hat{w} минимизируют функцию потерь

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

где

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

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 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 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

Поступила в редакцию