Сферические свёрточные нейронные сети для QSAR предсказаний*

Вареник Н. В., Попова М. С., Стрижов В. В.
varenik.nv@phystech.edu

Задача прогнозирования молекулярных свойств, например, биологической активности или растворимости на основе атомной структуры молекулы называется QSAR (Qualitative Structure Activity Relationships) предсказанием. Это классическая задача в области разработки лекарственных препаратов. Несмотря на то, что множество алгоритмов, таких как квантильная регрессия, нейронные сети на основе радиально-базисных функций являются приемлемыми решениями, все еще есть необходимость в более точной модели. В работе была выбрана модель сферических свёрточных нейронных сетей, изначально предложенная Taco S. Cohen et. al. для распознавания 3D-форм и положена под тщательное изучение в контексте QSAR предсказаний. Результаты исследуемой модели сравниваются с результатами более общих моделей, таких как реккурентные нейронные сети и случайный лес.

Ключевые слова: *QSAR предсказание, сферические свёрточные нейронные сети, разработка лекарств.*

1 Введение

Идея QSAR (Qualitative Structure Activity Relationships) заключается в нахождении связи между 2-х или 3-хмерным представлением молекулы и её биологическими или химическими свойствами. В связи с особенной важностью QSAR в сфере разрабатывания лекарственных препаратов в этой работе предлагается создать точный инструмент прогнозирования данной характеристики.

Изначально, было предложено использовать графическое представление молекулы для вычисления индекса Винера и терминального индекса Винера [6], которые коррелируют с такими понятиями как критическая точка [4], вязкость [3], но они не имеют четкой связи с растворимостью или активностью, которые особо важны в разработке лекарств.

Машинное обучение, как развивающаяся наука дает возможность используя ее различные методы, такие как случайный лес, квантильная и самосогласованная регрессии, нейронные сети постепенно улучшать качество прогнозирования в различных отрослях задачи нахождения QSAR.

Также рассмотривался вопрос рационального деления выборки на обучающую и тестовую [2]. Был сделан вывод, что оптимальный размер обучающей и тестовой выборки следует устанавливать на основе конкретного набора данных и типа используемых дескрипторов.

Стоит отметить модель нейронных сетей, предложенную в 2014 году и активно используемую в наши дни, так как она дает довольно неплохие результаты [5], в основе которой лежат радиальные базисные функции и самосогласованная регрессия.

В качестве решения рассматривается метод, основанный на сферических свёрточных нейронных сетях [1]. Они обладают уникальной особенностью, такой как возможность проектирования на плоскость сферического сигнала без искажений. Разработчик сферических CNN Taco et. al. протестировал их в различных задачах, в том числе в задаче

*Научный руководитель: Стрижов В. В. Задачу поставила: Попова М. С. Консультант: Попова М. С.

предсказания энергии распыления из молекулярной геометрии. Модель дала отличные результаты, поэтому возник интерес в её применении к задаче QSAR предсказаний. Основным недостатком предлагаемой модели является её сложность, связанная с большим числом параметров. Однако, ожидается, что данная модель станет универсальным решением нашей задачи. Результаты модели сравниваются с прогнозами более общих моделей, таких как RNN и случайный лес на данных, взятых MoleculeNet: A Benchmark for Molecular Machine Learning.

2 Постановка задачи

Пусть $\mathfrak{D} = (\mathbf{X}, \mathbf{y})$ — заданная выборка, где $\mathbf{X} \in \mathbb{R}^{m \times n \times 3}$ — тензор объект-признак, в нашем случае объекты $\mathbf{x}_i \in \mathbb{R}^{n \times 3}$ — это молекулы, каждая из которых описана вектором 3-хмерных координат всех ее атомов $\mathbf{x}_i = [x_i^1, \ldots, x_i^n]^\mathsf{T}, x_i^k \in \mathbb{R}^3, k = \overline{1,n}$, а $\mathbf{y} = [y_1, \ldots, y_m]^\mathsf{T} \in \mathbb{R}^m$ — свойства молекул. Их природа не имеет большого значения, это может быть растворимость, токсичность, биоактивность и т.п.

Рассмотрим множество параметрических моделей $\mathfrak{F}$, взятых из класса сферических сверточных нейронных сетей: $\mathfrak{F} = \{f_i \colon (\mathbf{w}, \mathbf{X}) \to \hat{\mathbf{y}} \mid i \in \mathfrak{I}\}$, где $\mathbf{w} \in \mathbf{W}$ - параметры модели, а $\hat{\mathbf{y}} \in \mathbb{R}^m$ - вектор предполагаемых свойств.

Задача состоит в предсказании свойства молекулы на основе её пространственной структуры. Будем рассматривать две задачи: задачу регрессии для предсказания численного значения определенного свойства и задачу классификации для предсказания наличия какого-нибудь свойства. Для регрессии считаем, что $\mathbf{y} \in N(\overline{\mathbf{y}}, \sigma_\mathbf{y})$, а для классификации $\mathbf{y} \in Be(p_\mathbf{y})$. Разобъем выборку на две части: обучающую ($\mathbf{X}_{\text{train}}$, $\mathbf{y}_{\text{train}}$) и тестовую ($\mathbf{X}_{\text{test}}$, $\mathbf{y}_{\text{test}}$).

Определим функцию ошибки:

– Для регрессии

$$S(\mathbf{y}, \mathbf{X}, \mathbf{w}) = \|\mathbf{y} - f(\mathbf{X}, \mathbf{w})\|_2^2 \tag{1}$$

– Для классификации

Параметры модели $\mathbf{w} \in \mathbf{W}$ подбираются в соответствии с минимизацией функции ошибки на обучении.

$$\mathbf{w}^* = \arg\min_{\mathbf{w} \in \mathbf{W}} S(\mathbf{y}_{\text{train}}, \mathbf{X}_{\text{train}}, \mathbf{w} \mid f) \tag{2}$$

Литература

[1] Taco Cohen, Mario Geiger, Jonas Koehler, and Max Welling. Spherical cnns. In *International Conference on Learning Representations*, March 2018.

[2] Alexander Golbraikh and Alexander Tropsha. Predictive qsar modeling based on diversity sampling of experimental datasets for the training and test set selection. *Journal of Computer-Aided Molecular Design*, 16(5-6):357–369, 2002.

[3] D. H. Rouvray and B. C. Crafford. The dependence of physical-chemical properties on topological factors. *South African Journal of Science*, 72:47, September 1976.

[4] Leonard I. Stiel and George Thodos. The normal boiling points and critical constants of saturated aliphatic hydrocarbons. *AIChE Journal*, 8:527–529, September 1962.

[5] Alexey V. Zakharov, Megan L. Peach, Markus Sitzmann, and Marc C. Nicklaus. A new approach to radial basis function approximation and its application to qsar. *Journal of Chemical Information and Modeling*, 54(3):713–719, 2014.

[6] Meryam Zeryouh, Mohamed El Marraki, and Mohamed Essalih. Some tools of qsar/qspr and drug development: Wiener and terminal wiener indices. In *Proceedings of 2015 International Conference on Cloud Computing Technologies and Applications (CloudTech?15)*, March 2015.