Сферические сверточные 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 сверточных 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 сверточные 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: ROC_AUC

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

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

3 Базовый эксперимент

3.1 Описание базовых алгоритмов

Цель проведения эксперимента - определить качество предсказания базовых моделей для дальнейшего сравнения результатов с исследуемой моделью.

Используемые данные состоят из 7165 молекул в формате SMILES, для которых предсказывалась энергия атомизации. Экспериментально было проверено, что зависимая переменная имеет нормальное распределение.

В качестве базовых моделей были взяты случайный лес и рекуррентная нейронная сеть. Случайный лес представляет собой ансамбль решающих деревьев, которые строятся по принципу жадной максимизации полученной информации: на каждом шаге алгоритм выбирает тот признак, который дает наибольший прирост информации при расщеплении, процедура повторяется рекурсивно до тех пор, пока энтропия не станет равной нулю

(или некоторому небольшому значению для учета переобучения). Оптимальные параметры для случайного леса были получены обычным перебором по сетке параметров и имеют значения n_estimators = 500, max_depth = 21.

Идея RNN заключается в последовательном использовании информации. То есть на каждом шаге скрытое состояние нейрона определяется как некоторая нелинейная функция от текущего входа и предыдущего сотояния нейрона. В эксперименте использовалась модель с параметрами:

- Число слоев в энкодере - 2
- Размер скрытого слоя - 128
- Число эпох - 70
- Ошибка - MSELoss
- Оптимизатор - Adam

	RMSE	R2_score
RNN	118.71	0.74
Random Forest	161.37	0.51

Таблица 1 Сравнение Random Forest и RNN

4 Сферическая свертка

Определение 1. *Единичная сфера* S *– это множество точек* $x \in \mathbb{R}^3 : ||x|| = 1$. *Это есть двумерное многообразие, которое может быть представлено с помощью сферических координат* $\alpha \in [0, 2\pi], \beta \in [0, \pi]$

Определение 2. *Сферический сигнал – это непрерывная функция* $f\colon S \to \mathbb{R}^K$, *где* K - *число каналов.*

Определение 3. *Группа вращений в 3-хмерном пространстве называется* $SO(3)$. *Это трехмерное многообразие, которое можно представить с помощью углов Эйлера* $\alpha \in [0, 2\pi], \beta \in [0, \pi], \gamma \in [0, 2\pi]$

Следует отметить, что результатом сферической свертки является не сигнал на сфере S, а а сигнал в $SO(3)$. Поэтому возникает необходимость определить свертку не только сферических сигналов, но и сигналов в $SO(3)$.

Определим понятие вращения сферического сигнала, чтобы знать, как вращать фильтр, который также является сферическим сигналом. Для этого введем оператор вращения:

$$[L_A f](x) = f(A^{-1}x) \tag{3}$$

где $x \in S$, $f\colon S \to \mathbb{R}^K$, – матрица вращения размера 3×3 ($||Ax|| = ||x||$ -сохраняет длину, $det(A) = +1$ - сохраняет ориентацию).

Произведение сферических сигналов $f, \varphi\colon S \to \mathbb{R}^K$ записывается следующим образом:

$$< \varphi, f >= \int_S \sum_{k=1}^{K} \varphi_k(x) f_k(x)\, dx \tag{4}$$

Тогда сферическая свертка двух сигналов f и φ определяется как:

$$[\varphi * f](A) =< L_A\varphi, f >= \int\limits_S \sum_{k=1}^{K} \varphi_k(A^{-1}x) f_k(x)\, dx \tag{5}$$

Теперь определим оператор вращения сигнала на $SO(3)$ аналогичным образом:

$$[L_A f](B) = f(A^{-1}B) \tag{6}$$

где $f\colon SO(3) \to \mathbb{R}^K$, , $B \in SO(3)$

Тогда свертка двух сигналов $f, \varphi\colon SO(3) \to \mathbb{R}^K$ на $SO(3)$ запишется как:

$$[\varphi * f](A) =< L_A\varphi, f >= \int\limits_{SO(3)} \sum_{k=1}^{K} \varphi_k(A^{-1}B) f_k(B)\, dB \tag{7}$$

где $A, B \in SO(3)$

Правомерность использования введенных выше сверток на всех слоях нейронной сети обосновывается важным свойсвом унитарности (сдвиг входа приводит к сдвигу выхода), т.е. слой Φ является унитарным, если $\Phi \circ L_A = T_A \circ \Phi$.

Для эффективного вычиления свертки используется обобщенный алгоритм быстрого преобразования Фурье (для функций, определенных на $S/SO(3)$), который есть ничто иное, как линейная проекция функции f на множество ортогональных базисных функций. Тогда обобщенное преобразование Фурье можно записать следующим образом:

$$\hat{f}^l = \int\limits_X f(x)\overline{U^l(x)}\, dx \tag{8}$$

где X - многообразие (S или $SO(3)$), $f\colon X \to \mathbb{R}$, $U^l(x)$ - базисные функции.

Определим также обратное $SO(3)$ преобразование Фурье:

$$f(A) = \sum_{l=0}^{b} (2l+1) \sum_{m=-l}^{l} \sum_{n=-l}^{l} \hat{f}^l_{mn} D^l_{mn}(A) \tag{9}$$

где $D^l_{mn}(A)$ - винеровская D-функция, $l \geqslant 0, -l \leqslant m, n \leqslant l$

Для функций на сфере вместо D^l_{mn} используется сферическая гармоника $Y^l_m = D^l_{m0}|_S, l \geqslant 0, -l \leqslant m \leqslant l$.

Оказывается, для сферических сигналов и для сигналов на $SO(3)$ выполнена теорема Фурье: $\hat{\varphi * f} = \hat{f} \cdot \hat{\varphi}^\dagger$, где "$\cdot$"обозначает матричное произведение, а в случае Фурье-преобразования сферического сигнала теорема принимает вид: $\hat{\varphi * f^l} = \hat{f}^l \cdot \hat{\varphi}^{l\dagger}$, только здесь уже "$\cdot$ поэлементное перемножение векторов.

Теперь стало возможным описание полноценного слоя свертки: сигнал f и фильтр φ преобразовываются обобщенным быстрым преобразованием Фурье, затем они перемножаются. Получаем блочный тензор, который суммируется по всем каналам, а после этого подвергается обратному $SO(3)$ преобразованию Фурье.

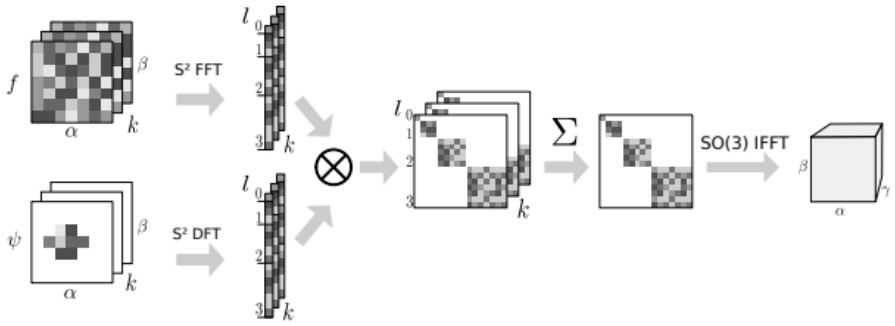

Рис. 1 Слой сферической свертки: сферический сигнал параметризован сферическими координатами α, β, а сигнал на SO(3) - углами Эйлера α, β, γ

5 Вычислительный эксперимент (Прогнозирование энергии атомизации)

Молекулы состоят из $N = 23$ атомов $T = 5$ типов (H, C, N, O, S). Они даны в виде списков позиций p_i и зарядов z_i каждого атома i.

Представление молекул кулоновскими матрицами Кулоновская матрица $\mathbf{C} \in \mathbb{R}^{N \times N}$ – это инвариантное относительно перемещения и вращения представление молекул. Для каждой пары атомов $i \neq j$ определяется $C_{ij} = (z_i z_j)/(|p_i - p_j|)$, а $C_{ii} = 0.5 z_i^{2.4}$. Диагональные элементы кодируют атомную энергию, образованную зарядом ядра, а другие элементы матрицы кодируют кулоновское отталкивание между атомами.

TODO: описать построение сферического сигнала, архитектуру сети

	RMSE	R2_score
Random Forest	161.37	0.51
RNN	118.71	0.74
SCNN	5.51	

Таблица 2 Результаты

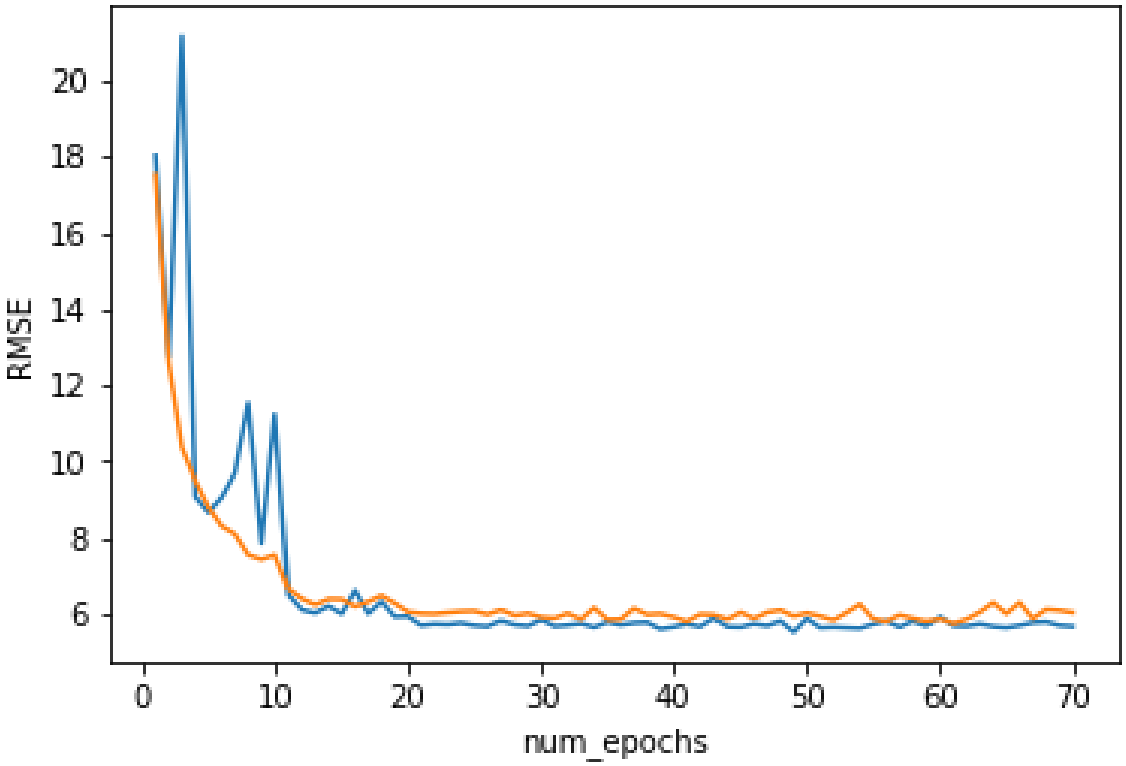

Рис. 2 Результаты Scnn

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.