Сферические сверточные нейронные сети для прогнозирования молекулярных свойств*

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

Рассматривается задача прогнозирования молекулярных свойств на основе атомной структуры молекулы, называемая QSPR (Qualitative Structure Property Relationships) прогнозированием. Это классическая задача в области разработки лекарственных препаратов. Несмотря на то, что множество методов таких, как использование SVM, случайного леса или рекуррентной нейронной сети являются приемлемыми решениями, все еще есть необходимость в более точной модели. В работе выбрана модель сферических сверточных нейронных сетей, изначально предложенная Taco S. Cohen et. al. для распознавания 3D-форм и положена под тщательное изучение в контексте рассматриваемой задачи. Результаты исследуемой модели сравниваются с результатами реккурентной нейронной сети и случайного леса.

Ключевые слова: *QSPR прогнозирование, сферические сверточные нейронные сети, разработка лекарств.*

1 Введение

Количественная взаимосвязь структура-свойство (QSPR) заключается в построении модели прогнозирования свойства молекулы, как функции от ее структуры [5]. Под структурой подразумевается 2D/3D представление молекулы. Подход QSPR широко используется в области разработки лекарственных препаратов [1]. В процессе разработки существенным является определение влияния структурных особенностей в молекуле лекарства на различные свойства, поскольку это позволяет получить информацию о важных функциональных группах в структурах тестируемых соединений. Таким образом, меняя некоторые группы в структурах препаратов можно повысить их фармакологическую активность или биологическо-химические свойсва. Экспериментальное определение взаимосвязи структуры со свойством является дорогостоящей процедурой, а QSPR позволяет снизить эту цену. В связи с этим в работе предлагается новый метод прогнозирования данной характеристики.

Среди существующих методов для решения задачи часто используется модель случайного леса с различными методами оптимизации параметров. В [7] рассматривается алгоритм поиска оптимального числа деревьев. Структура молекулы в большинстве случаев представляется в виде последовательности SMILES [6], поэтому рекуррентная нейронная сеть [4] нередко используется в области QSPR прогнозирования. В [3] можно убедиться в преимуществах использования различных нейронных сетей, в том числе реккурентной нейронной сети для предсказания QSPR.

В данной работе предлагается рассмотреть модель сферических свёрточных нейронных сетей(SCNN) [2], которая отличается от обычной CNN тем, что операция свертки осуществляется над сферическим сигналом с фильтром, который также является сферическим сигналом. Еще одно отличие состоит в том, что SCNN обладает возможностью проектирования сферического сигнала на плоскость без искажений. Основным недостатком предлагаемой модели является её сложность, связанная с большим числом параметров.

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

Так как, используемые на данный момент модели прогнозирования не учитывают взаимного расположения атомов в пространстве в отличии SCNN предполагается, что качество прогноза будет улучшено. Результаты исследуемой модели сравниваются с вышеупомянутыми моделями RNN и случайного леса на данных, взятых из MoleculeNet: A Benchmark for Molecular Machine Learning.

2 Постановка задачи

Пусть $\mathfrak{D} = (\mathbf{X}, \mathbf{y})$ — заданная выборка, где $\mathbf{X} \in \mathbb{R}^{m \times n \times 4}$ — тензор объект-признак, объекты $\mathbf{x}_i \in \mathbb{R}^{1 \times n \times 4}, i = \overline{1, m}$ — это молекулы, каждая из которых описана множеством 3-мерных координат всех ее атомов и зарядом, а $\mathbf{y} = [y_1, \ldots, y_m]^\mathsf{T} \in \mathbb{R}^{m \times 1}$ — 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: $\mathfrak{F} = \{\mathbf{f}_k \colon (\mathbf{w}, \mathbf{X}) \to \hat{\mathbf{y}} \mid k \in \mathfrak{K}\}$, где $\mathbf{w} \in \mathbb{W}$ – параметры модели, а $\hat{\mathbf{y}} \in \mathbb{R}^{m \times 1}$ - вектор предполагаемых свойств.

Задача состоит в прогнозировании свойства y_i молекулы на основе её пространственной структуры $\mathbf{x}_i$. Рассматривается задача регрессии для предсказания численного значения определенного свойства. Считаем, что $\mathbf{y} \in \mathcal{N}(\mathsf{E}(\mathbf{y}), \sigma_\mathbf{y})$.

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: $S(\mathbf{y}, \mathbf{X}, \mathbf{w}) = \|\mathbf{y} - \mathbf{f}(\mathbf{X}, \mathbf{w})\|_2$

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

$$\mathbf{w}^* = \arg\min_{\mathbf{w} \in \mathbb{W}} S(\mathbf{w} \mid \mathbf{f}, \mathbf{y}_{\text{train}}, \mathbf{X}_{\text{train}}) \tag{1}$$

3 Базовый эксперимент(прогнозирование энергии атомизации)

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 Процесс сферической свертки

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. Всего 7165 молекул.

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.

Построение сферического сигнала Вокруг каждого ненулевого атома i строится сфера S_i постоянного радиуса и такого, чтобы не было перечесения в обучающем наборе. Затем для каждого уникального z и для каждого $x \in S_i$ определяется потенциальная функция $U_z(x) = \sum_{j\neq i, z_j = z} \frac{z_i \cdot z}{|x-p_i|}$ производящая сферический сигнал. После этого сигнал дискретизируется проектированием на сетку Driscoll-Healy (Driscoll and Healy, 1994) с шириной полосы $b = 10$.

Определение 1. *Единичной сферой S называется множество точек $\mathbf{x} \in \mathbb{R}^3 : ||\mathbf{x}|| = 1$. 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 $\mathbf{g}\colon S \to \mathbb{R}^K$, где K - число каналов.*

Определение 3. *$SO(3)$ – группа вращений в 3-мерном пространстве . 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:

$$[L_A f](x) = f(A^{-1}x) \tag{2}$$

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:

$$\langle \varphi, f \rangle = \int\limits_S \sum_{k=1}^{K} \varphi_k(x) f_k(x)\, dx \tag{3}$$

Тогда сферическая свертка двух сигналов f и φ определяется как:

$$[\varphi * f](A) = \langle L_A \varphi, f \rangle = \int\limits_S \sum_{k=1}^{K} \varphi_k(A^{-1}x) f_k(x)\, dx \tag{4}$$

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

$$[L_A f](B) = f(A^{-1}B) \tag{5}$$

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:

$$[\varphi * f](A) = \langle L_A \varphi, f \rangle = \int\limits_{SO(3)} \sum_{k=1}^{K} \varphi_k(A^{-1}B) f_k(B)\, dB \tag{6}$$

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:

$$\hat{f}^l = \int\limits_X f(x) \overline{U^l(x)}\, dx \tag{7}$$

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:

$$f(A) = \sum_{l=0}^{b} (2l+1) \sum_{m=-l}^{l} \sum_{n=-l}^{l} \hat{f}^l_{mn} D^l_{mn}(A) \tag{8}$$

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: $\widehat{\varphi * f} = \hat{f} \cdot \hat{\varphi}^\dagger$, где " $\cdot$ " обозначает матричное произведение, а в случае Фурье-преобразования сферического сигнала теорема принимает вид: $\widehat{\varphi * f^l} = \hat{f}^l \cdot \hat{\varphi}^{l\dagger}$, 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.

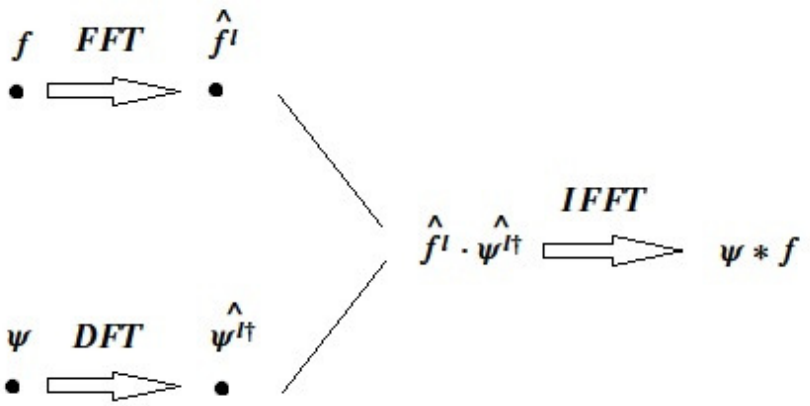

Рис. 1 Слой сферической свертки

5 Вычислительный эксперимент (Прогнозирование энергии атомизации)

Цель эксперимента - определить качество предсказания модели сферических сверточных нейронных сетей. Используемые данные описаны в разделе 4.

В отличии от обычной сверточной нейронной сети, где фильтр перемещается по проскости изображения, в сверической - фильтр вращается вокруг поверхности единичной сферы и выделяет определенные признаки уже из поверхности сферы. Основной особенностью сферической свертки является эквивариантность относительно вращений, то есть мы будем получать одинаковые ответы даже если исследуемый объект будет вращаться. Для уменьшения вычислительной сложности будем преобразовывать сферические сигналы с помощью преобразования Фурье и тогда для свертки нужно будет просто перемножить преобразованные сигналы, а затем воспользоваться обратным преоразованием Фурье для получения результата свертки сферических сигналов.

Параметры сети:

- Число эпох – 70
- Ошибка – MSELoss
- Оптимизатор – Adam
- Размер батча – 32

Результаты эксперимента:

	RMSE	R2_score
Random Forest	161.37	0.51
RNN	118.71	0.74
SCNN	5.51	0.99

Таблица 2 Сравнение SCNN с RNN и RF

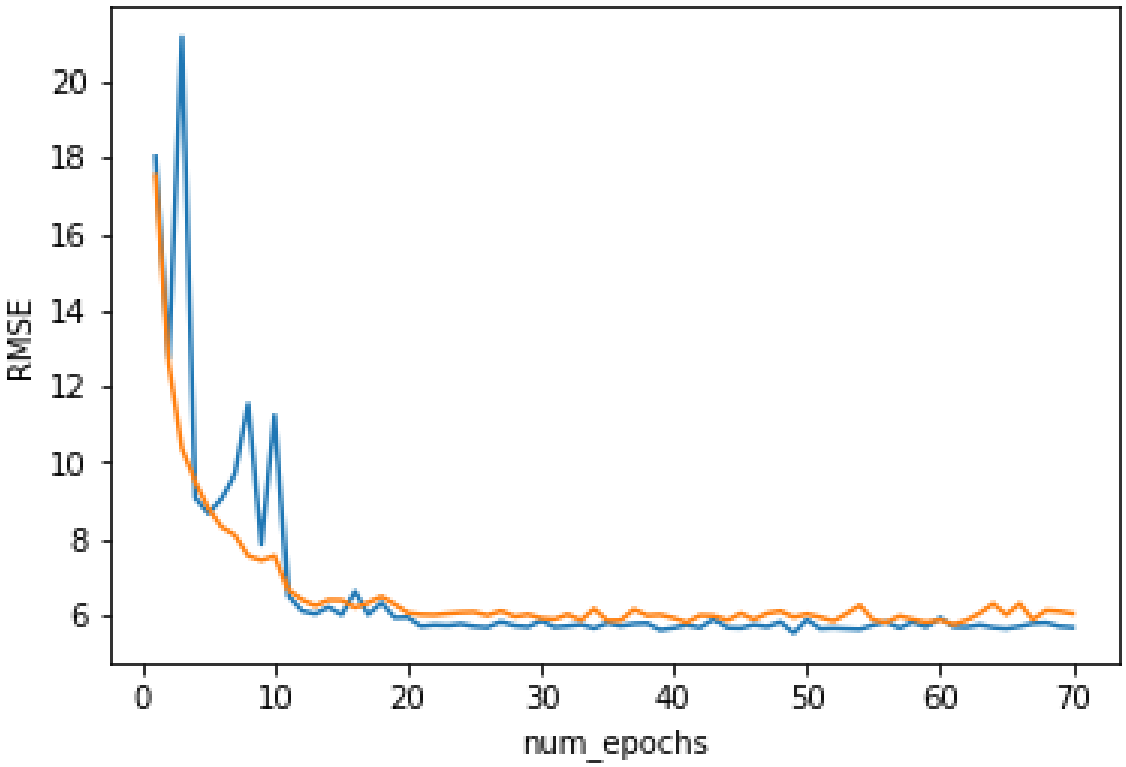

Рис. 2 Результаты Scnn

Литература

[1] Ajita Atul Bhat. *Applications of bioinformatics, homology modeling and QSAR to drug design.* PhD thesis, University of Georgia, 1997.

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

[3] Garrett B. Goh, Nathan O. Hodas, and Abhinav Vishnu. Deep learning for computational chemistry. *Journal of Computational Chemistry*, 38(16):1291–1307, 2017.

[4] Esther Levin. A recurrent neural network: Limitations and training. *Neural Networks*, 3(6):641–650, 1990.

[5] Chanin Nantasenamat, Chartchalerm Isarankura-Na-Ayudhya, Thanakorn Naenna, and Virapong Prachayasittikul. A practical overview of quantitative structure-activity relationship. 2003.

[6] Noel M. O'Boyle. Towards a universal smiles representation - a standard method to generate canonical smiles based on the inchi. *J. Cheminformatics*, 4:22, 2012.

[7] Pavel G. Polishchuk, Eugene N. Muratov, Anatoly G. Artemenko, Oleg G. Kolumbin, Nail N. Muratov, and Victor Kuzmin. Application of random forest approach to qsar prediction of aquatic toxicity. *Journal of Chemical Information and Modeling*, 49(11):2481–2488, 2009.