Сферические сверточные нейронные сети для прогнозирования химических свойств молекул

Наталия Викторовна Вареник

Московский физико-технический институт

Курс: Численные методы обучения по прецедентам (практика, В. В. Стрижов)/Группа 674, весна 2019

Прогнозирование химических свойств

Цель исследования

Повысить качество прогнозирования химических свойств путем учитывания 3-мерной пространственной структуры молекулы.

Проблема

Существующие методы ввиду своего построения могут учитывают лишь плоскую структуру молекулы.

Метод решения

Предлагается использовать модель сферических сверточных нейронных сетей, учитывающую пространственное представление исследуемого объекта через сферический сигнал, построенный по признаковому описанию.

Литература

Модели, не учитывающие пространственное представление:

- David S. Palmer and Noel M. O'Boyle and Robert C. Glen and John B. O. Mitchell. *Random Forest Models To Predict Aqueous Solubility*, 2007.
- George E. Dahl and Navdeep Jaitly and Ruslan Salakhutdinov. *Multi-task Neural Networks for QSAR Predictions*, 2014.

Модель сферических сверточных нейронных сетей:

- Taco S. Cohen, Mario Geiger, Jonas Koehler, Max Welling. *Spherical CNNs*, 2018.

Постановка задачи

Дано

Выборка $\mathfrak{D} = (\mathbf{X}, \mathbf{y})$, где $\mathbf{X} \in \mathbb{R}^{m \times n \times 4}$ — тензор объект-признак, состоящий из m молекул, в каждой из которых по n атомов, описанных 3-мя координатами и зарядом, а $\mathbf{y} \in \mathbb{R}^m$ — свойства молекул.

Модель

$\mathfrak{F} \ni \mathbf{f} : (\mathbf{w}, \mathbf{X}) \to \hat{\mathbf{y}}$, где $\mathbf{w} \in \mathbb{W}$ – параметры модели, а $\mathfrak{F}$ - семейство параметрических моделей из класса сферических сверточных нейронных сетей.

Критерий качества

- функция ошибки:

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

- дополнительная метрика качества:

$$R2 = 1 - \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{m}(y_i - \overline{y}_i)^2}$$

Итоговая задача оптимизации

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

$$\mathbf{w}^* = \underset{\mathbf{w} \in \mathbb{W}}{\operatorname{argmin}} S(\mathbf{w} \mid f, \mathbf{y}_{\text{train}}, \mathbf{X}_{\text{train}})$$

Построение сферического сигнала

- Вокруг каждого ненулевого атома i строится сфера S_i постоянного радиуса и такого, чтобы не было перечесения в обучающем наборе
- 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.
- Сигнал дискретизируется проектированием на сетку Driscoll-Healy (Driscoll and Healy, 1994) с шириной полосы $b = 10$.

Сферическая свертка

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

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

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

Тогда сферическая свертка двух сигналов f и ϕ определяется как:

$$[\phi * f](A) = \langle L_A \phi, f \rangle = \int_S \sum_{k=1}^{K} \phi_k(A^{-1}x) f_k(x)\, dx \tag{2}$$

Вычислительный эксперимент(базовые модели)

Данные

Формат представления молекулы – SMILES,
число молекул – 7165, число атомов – 23,
целевая переменная – энергия атомизации.

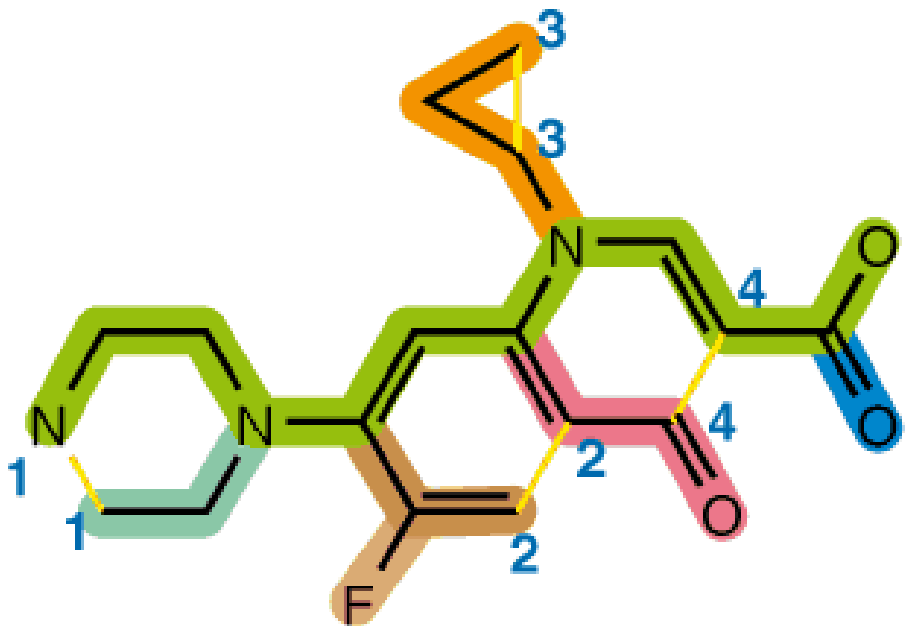
Модель Random Forest

n_estimators – 500, max_depth – 21

Модель RNN

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

Пример построения последовательности SMILES



N1CCN(CC1)C(C(F)=C2)=CC(=C2C4=O)N(C3CC3)C=C4C(=O)O

Brian Satola. Convert CAS Registry Number to Other Identifiers, 2017

Вычислительный эксперимент(основная модель)

Цель

Определить качество предсказания модели сферических сверточных нейронных сетей.

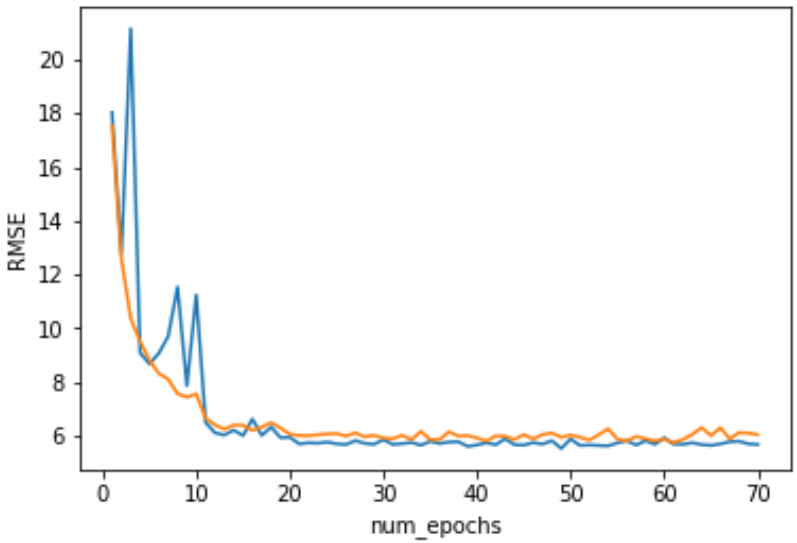
Данные

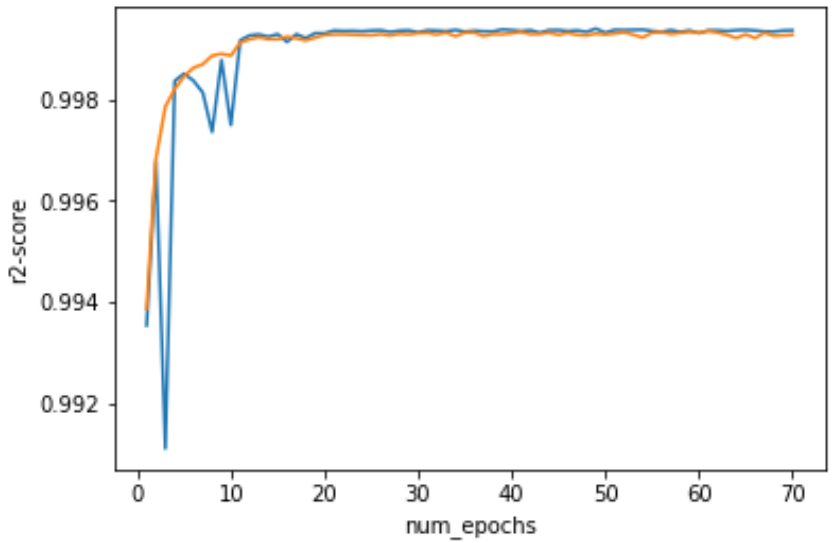
Молекула представлена множеством зарядов и 3-мерных координат ее атомов, число молекул - 7165, число атомов - 23, целевая переменная - энергия атомизации.

Модель SCNN

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

Результаты





	RMSE	R2-score
RF	161.37	0.51
RNN	118.71	0.74
SCNN	5.51	0.99

Выводы

- Предложен новый метод прогнозирования свойств молекул
- Определен способ построения сферического сигнала для исследуемого класса задач
- Продемонстрировано значительное улучшение качества прогноза предложенного метода по сравнению с базовыми, которые чаще всего применяются на данный момент.