

Решение задачи оптимизации, сочетающей классификацию и регрессию, в применении к молекулярному докингу

Антон Юрьевич Бишук

Московский физико-технический институт

Курс: Название нашего курса/Группа 774, весна 2020

Консультанты: Сергей Грудинин, Мария Кадукова

Цель работы

Определение свободной энергии связывания для комплексов «белок-лиганд», которая наилучшим образом предсказывает нативные конформации и при этом имеет высокую корреляцию с энергией связывания, полученной экспериментально.

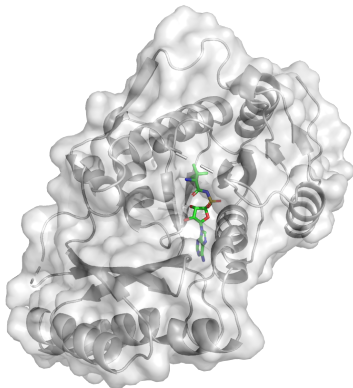
Существующие проблемы

- В силу высокой размерности признакового пространства, задача является вычислительно сложной;
- Существующие решения нахождения нативных конформаций, дают слабую корреляцию с энергией связывания.

Способ решения

Объединение существующих методов классификации конформаций и регрессии в одну оптимизационную задачу.

Постановка задачи: основные термины и понятия



Комплекс «белок-лиганд»

- Белок (лиганд) – граф, вершины которого соответствуют атомам, а ребра – ковалентным связям между ними;
- Комплекс – система из точек белка и лиганда;
- Конформация – взаимное расположение белка и лиганда.

Постановка задачи: модель взаимодействия

Пусть имеется M_1 типов атомов белка и M_2 типов атомов лиганда. Возьмем в качестве скорингового функционала свободную энергию связывания E в следующих предположениях:

- E определяется только взаимодействиями между парами атомов рассматриваемого комплекса;
- E зависит только от распределения расстояний между взаимодействующими атомами;
- E является линейным функционалом.

Тогда он имеет следующий вид:

$$E(n(r)) = \sum_{k=1}^{M_1} \sum_{l=1}^{M_2} \int_0^{r_{\max}} n^{kl}(r) f^{kl}(r) dr$$

- $n^{kl}(r) = \sum_{i,j} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(r-r_{ij})^2}{2\sigma^2}\right]$
- $f^{kl}(r)$ – неизвестные функции взаимодействия между атомами типов k и l .

После разложения функций $n^{kl}(r)$ и $f^{kl}(r)$ по полиномиальному базису, $E(n(r))$ принимает вид:

$$E(n(r)) \approx \sum_{k=1}^{M_1} \sum_{l=1}^{M_2} \sum_{q=0}^Q w_q^{kl} x_q^{kl} = \langle \mathbf{w}, \mathbf{x} \rangle,$$

$$\mathbf{w}, \mathbf{x} \in \mathbb{R}^{Q \times M_1 \times M_2}$$

- Q – порядок разложения;
- Вектор \mathbf{w} будем называть *скоринговым вектором*;
- Вектор \mathbf{x} – *структурным вектором*.

Постановка задачи: оптимизационная задача

Дано

Множество троек $\{\mathbf{x}, y_i, s_i\}_i^N$, где \mathbf{x} – структурный вектор,
 $y_i \in \{-1, 1\}$ – поза i -го комплекса (нативной конформации соответствует значение $y_i = 1$, ненативной $y_i = -1$)
 s_i – значение энергии связывания i -го комплекса.

Общий вид задачи оптимизации

$$\begin{aligned} \underset{\mathbf{w}, b_i, \xi_i}{\text{Minimize:}} \quad & \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_i \xi_i + C_r \sum_i f^{nat}(\mathbf{X}_i, \mathbf{w}, s_i) \\ & y_i [\mathbf{w}^T \mathbf{X}_i - b_i] - 1 + \xi_i \geq 0 \\ \text{Subject to:} \quad & \xi_i \geq 0 \end{aligned}$$

- $f^{nat}(\mathbf{X}_i, \mathbf{w}, s_i)$ – функция потерь регрессии,
- C_r, C – коэффициент регуляризации для функции потерь регрессии и классификации соответственно.

Постановка задачи: подробности

Взяв в качестве функции потерь $f(\mathbf{X}_i, \mathbf{w}, s_i) = (\langle \mathbf{X}_i, \mathbf{w} \rangle - s_i)^2$ и упростив выражение, введем замену переменных $\mathbf{w}' = \mathbf{A}\mathbf{w} - \mathbf{d}$,

$$\mathbf{A} = \left[\mathbf{I} + 2C_r \mathbf{X}^{nat T} \mathbf{X}^{nat} \right]^{\frac{1}{2}} \quad \mathbf{d} = 2C_r \left[\mathbf{I} + 2C_r \mathbf{X}^{nat T} \mathbf{X}^{nat} \right]^{-\frac{1}{2}} \mathbf{X}^{nat T} \mathbf{s}$$

Задача оптимизации примет вид

$$\begin{aligned} \text{Minimize:} \quad & \frac{1}{2} \|\mathbf{w}'\|^2 + C \sum_i \xi_i \\ & y_i [(\mathbf{A}^{-1}(\mathbf{w}' + \mathbf{d}))^T \mathbf{X}_i - b_i] - 1 + \xi_i \geq 0 \\ \text{Subject to:} \quad & \xi_i \geq 0 \end{aligned}$$

Двойственная задача

$$\begin{aligned} \text{Minimize:} \quad & \frac{1}{2} \sum_{(i,j)=1}^{P \cdot D} \lambda_i \lambda_j y_i y_j \langle \hat{\mathbf{X}}_i, \hat{\mathbf{X}}_j \rangle + \sum_{i=1}^{P \cdot D} \lambda_i (y_i \langle \mathbf{d}, \hat{\mathbf{X}}_i \rangle - 1) \\ \text{Subject to:} \quad & 0 \leq \lambda_i \leq C, \quad \sum_{j=1}^D \lambda_{k+j} y_{k+j} = 0, \quad k \in \{0, D, 2D, \dots, D(P-1)\}. \end{aligned}$$

Вычислительный эксперимент: данные

Данные представляют из себя размеченную матрицу признаков для 12,000 комплексов белков с лигандами, для каждого из которых известна одна нативная и 18 ненативных конформаций.

- **Данные для бинарной классификации:**

- Основными признаками объектов являются гистограммы распределений расстояний между парами атомов белка и лиганда различных типов.
- Так же в данных есть вектор меток y , у которого на месте нативной конформации стоит единица, а на месте ненативной конформации -1.

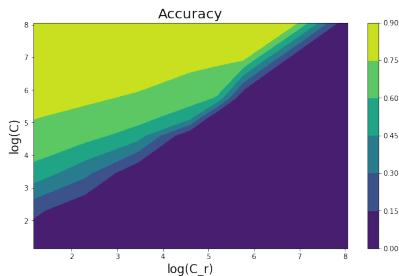
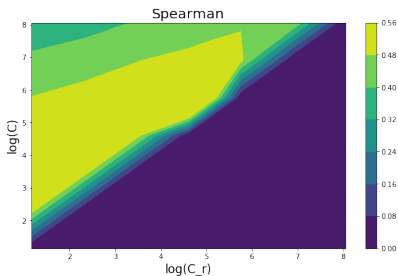
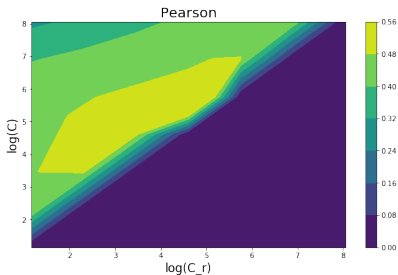
- **Данные для регрессии:**

- Дополнительные 379 признаков мы используем для учета взаимодействий комплексов с растворителем, а так же энтропийных потерь из-за ограничения подвижности лиганда.
- Для каждого из представленных комплексов известно значение величины, которую можно интерпретировать как энергию связывания.

Вычислительный эксперимент: План работы

- Признак, отвечающий за подвижность атома одинаковый для всех конформаций одного комплекса, поэтому первым делом необходимо проверить, что его добавление не ухудшает классификацию;
- Проверить работу алгоритма, если использовать все имеющиеся признаки;
- Провести подбор оптимальных параметров модели на кросс-валидации;
- Провести тесты на докинг-тестах из бенчмарка CASF.

Вычислительный эксперимент: нахождение оптимальных параметров на сетке



На данный момент:

- Сформулирована оптимизационная задача;
- Проведен эксперимент на выборке, включающей в себя 1000 комплексов
- Произведен грубый подбор оптимальных параметров.

Предстоит:

- Произвести отбор признаков на основе физической значимости и математической важности для модели;
- Проведен эксперимент докинг-тестах;
- Оценить необходимые ресурсы для того, чтобы обучить модель на всех данных;
- Подобрать более оптимальные параметры, путем уменьшения шага солвера и сетки.