Формулировка и решение задачи оптимизации, сочетающей классификацию и регрессию, для оценки энергии связывания белка и маленьких молекул

Анастасия Грачева, Мария Кадукова, Сергей Грудинин, В.В. Стрижов
gracheva.as@phystech.edu
[1]ФИВТ МФТИ

При разработке лекарства возникает задача поиска маленьких молекул - лигандов, наиболее сильно взаимодействующих с исследуемым белком, а значит являющихся основными кандидатами в лекарства. Один из способов определить действительное положение такого лиганда заключается в том, чтобы генерировать несколько возможных положений и классифицировать их как нативные и не нативные. Но качество предсказания может быть повышено, если использовать экспериментальные данные о свободной энергии связывания молекул и решать одновременно задачи регрессии и классификации. В статье будут рассмотрены эксперименты с алгоритмом, использующим эту идею.

1 Введение

Предсказание наиболее выгодной ориентации и положения молекул по отношению друг к другу для образования устойчивого комплекса из белка и лиганда, или молекулярный докинг - задача, важная для ускорения процесса разработки новых лекарств. Есть два метода её решения: pose prediction - среди нескольких сгенерированных положений лиганда в белке определить наиболее близкое к реальному и scoring - предсказать аффинность (свободную энергию связывания) для комплексов различных белков с лигандами. При этом положение с наименьшей энергией связывания будет соответствовать нативной конформации. Первая задача решена в работе [1] с помощью оптимизации скоринговой функции, учитывающей всевозможные комбинации различных пар атомов и расстояния между ними. Раскладывая эту функцию по базису, авторы представляют её как вектор структурных коэффициентов и сводят задачу к модифицированной SVM-классификации.

Наше предположение заключается в том, что с использованием экспериментальных данных об аффинностях можно улучшить качество классификации. В эксперименте, описанном в данной статье, мы проверим эту гипотезу, а также постараемся решать оптимизационную задачу максимально эффективно вычислительно, чтобы использовать как можно больше доступных экспериментальных данных.

2 Постановка задачи

Пусть $\{C_{ij}\}_{i=1}^{P}$ - комплексы белков и лигандов. При $j = 0$ они находятся в нативных позах, при $j = 1 \dots D$ - в ненативных. Задача заключается в том, чтобы найти скоринговую функцию E, который удовлетворяет неравенствам:

$$\begin{aligned} &E(C_{i0}) < E(C_{ij}) \\ &\forall i \in 1, \dots, P, \\ &\forall j \in 1, \dots, D. \end{aligned} \tag{1}$$

В модели взаимодействия, описанной в [1], эта функция задаётся скоринговым вектором $\mathbf{w}$. Поэтому неравенство выше может быть преобразовано в систему неравенств:

$$\begin{aligned}
&\langle \mathbf{x}_{i0}, \mathbf{w} \rangle < \langle \mathbf{x}_{ij}, \mathbf{w} \rangle, \\
&\langle \mathbf{x}_{ij} - \mathbf{x}_{i0}, \mathbf{w} \rangle > 0, \\
&\forall i = 1, \ldots, P, \\
&\forall j = 1, \ldots, D.
\end{aligned} \tag{2}$$

где x_{ij} - структурные вектора.

Чтобы гарантировать единственность решения, а также решать задачу в случае линейной неразделимости выборки, приведём её к виду задачи квадратичной оптимизации с мягким зазором:

$$\begin{aligned}
\underset{\mathbf{w}, b_i, \xi_{ij}}{\text{minimize:}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{ij} \xi_{ij} \\
\text{subject to:} \quad & y_{ij}[\langle \mathbf{w}, \mathbf{x}_{ij} \rangle - b_i] - 1 + \xi_{ij} \geqslant 0, \\
& \xi_{ij} \geqslant 0, \\
& i \in \{1, \ldots, P\}, \\
& j \in \{0, \ldots, D\},
\end{aligned} \tag{3}$$

где $\mathbf{w}$, b_i и переменные невязки ξ_{ij} – оптимизируемые параметры модели,
$y_{i0} = 1$ для нативной позы и $y_{ij} = -1,\ j \in \{1, \ldots, D\}$, для ненативной,
а C – некоторый коэффициент регуляризации.
Таким образом, решив оптимизационную задачу, решаем и задачу класификации.

Кроме того, есть другая потановка этой задачи - задача регрессии, т.е. предсказания значения свободной энергии связывания белка с лигандом:

$$\begin{aligned}
\underset{\mathbf{w}}{\text{minimize:}} \quad & \sum_i [\langle \mathbf{w}, \mathbf{x}_{i0} \rangle - s_i]^2 + \alpha\|\mathbf{w}\|^2, \\
& i \in \{1, \ldots, P\},
\end{aligned} \tag{4}$$

где s_i – экспериментально полученное значение энергии связывания i-го нативного соединения, α – коэффициент регуляризации для ridge-регрессии.

Объединение этих методов заключается в сложении функций потерь классификации и регрессии:

$$\begin{aligned}
\underset{\mathbf{w}, b_i, \xi_{ij}}{\text{minimize:}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{ij} \xi_{ij} + C_r \sum_i f(\mathbf{x}_{i0}, \mathbf{w}, s_i) \\
\text{subject to:} \quad & y_{ij}[\langle \mathbf{w}, \mathbf{x}_{ij} \rangle - b_i] - 1 + \xi_{ij} \geqslant 0, \\
& \xi_{ij} \geqslant 0, \\
& i \in \{1, \ldots, P\}, \\
& j \in \{0, \ldots, D\},
\end{aligned} \tag{5}$$

где $f(\mathbf{x}_{i0}, \mathbf{w}, s_i)$ – MSE, C_r – коэффициент регуляризации для функции потерь регрессии.

3 Теоретическая часть

С помощью замены переменных сведем два квадратичных слагаемых в целевой функции из задачи (5) к одному.

Функция потерь регрессии MSE для одного комплекса выражается формулой:

$$f(\mathbf{x}_{i0}, \mathbf{w}, s_i) = (\langle \mathbf{x}_{i0}, \mathbf{w} \rangle - s_i)^2. \tag{6}$$

Тогда для выборки $\mathbf{X} = (\mathbf{x}_{10}, \ldots, \mathbf{x}_{P0})^{\mathrm{T}}$, состоящей из нативных конфигураций, и целевого вектора $\mathbf{s} = (s_1, \ldots, s_P)^{\mathrm{T}}$ квадратичные слагаемые целевой функции из задачи (5) принимают вид:

$$
\begin{aligned}
&\frac{1}{2}\|\mathbf{w}\|^2 + C_r \sum_i f(\mathbf{x}_{i0}, \mathbf{w}, s_i) = \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + C_r\|\mathbf{X}\mathbf{w} - \mathbf{s}\|^2 = \\
&= \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + C_r\mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w} - 2C_r\mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{s} + C_r\mathbf{s}^{\mathrm{T}}\mathbf{s} = \\
&= \mathbf{w}^{\mathrm{T}}\left(\frac{1}{2}\mathbf{I} + C_r\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)\mathbf{w} - 2C_r\mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{s} + C_r\mathbf{s}^{\mathrm{T}}\mathbf{s} = \\
&= \left\|\left[\frac{1}{2}\mathbf{I} + C_r\mathbf{X}^{\mathrm{T}}\mathbf{X}\right]^{\frac{1}{2}}\mathbf{w}\right\|^2 - 2C_r\left(\mathbf{w}^{\mathrm{T}}\left[\frac{1}{2}\mathbf{I} + C_r\mathbf{X}^{\mathrm{T}}\mathbf{X}\right]^{\frac{1}{2}}\right)\left(\left[\frac{1}{2}\mathbf{I} + C_r\mathbf{X}^{\mathrm{T}}\mathbf{X}\right]^{-\frac{1}{2}}\mathbf{X}^{\mathrm{T}}\mathbf{s}\right) + C_r\mathbf{s}^{\mathrm{T}}\mathbf{s} = \\
&= \left(\left[\frac{1}{2}\mathbf{I} + C_r\mathbf{X}^{\mathrm{T}}\mathbf{X}\right]^{\frac{1}{2}}\mathbf{w} - C_r\left[\frac{1}{2}\mathbf{I} + C_r\mathbf{X}^{\mathrm{T}}\mathbf{X}\right]^{-\frac{1}{2}}\mathbf{X}^{\mathrm{T}}\mathbf{s}\right)^2 + C_r\mathbf{s}^{\mathrm{T}}\mathbf{s} - C_r^2\left\|\left[\frac{1}{2}\mathbf{I} + C_r\mathbf{X}^{\mathrm{T}}\mathbf{X}\right]^{-\frac{1}{2}}\mathbf{X}^{\mathrm{T}}\mathbf{s}\right\|^2.
\end{aligned}
$$

Введем замену переменных:

$$
\begin{aligned}
&\mathbf{w}' = \mathbf{A}\mathbf{w} - \mathbf{B}, \text{ где} \\
&\mathbf{A} = \left[\frac{1}{2}\mathbf{I} + C_r\mathbf{X}^{\mathrm{T}}\mathbf{X}\right]^{\frac{1}{2}}, \\
&\mathbf{B} = C_r\left[\frac{1}{2}\mathbf{I} + C_r\mathbf{X}^{\mathrm{T}}\mathbf{X}\right]^{-\frac{1}{2}}\mathbf{X}^{\mathrm{T}}\mathbf{s}.
\end{aligned}
\tag{7}
$$

Тогда, учитывая, что

$$
C_r\mathbf{s}^{\mathrm{T}}\mathbf{s} - C_r^2\left\|\left[\frac{1}{2}\mathbf{I} + C_r\mathbf{X}^{\mathrm{T}}\mathbf{X}\right]^{-\frac{1}{2}}\mathbf{X}^{\mathrm{T}}\mathbf{s}\right\|^2 = \mathrm{const},
$$

задача оптимизации принимает вид:

$$
\begin{aligned}
\underset{\mathbf{w}', b_i, \xi_{ij}}{\text{minimize:}} \quad & \frac{1}{2}\|\mathbf{w}'\|^2 + C\sum_{ij}\xi_{ij} \\
\text{subject to:} \quad & y_{ij}[\left(\mathbf{A}^{-1}(\mathbf{w}' + \mathbf{B})\right)^{\mathrm{T}}\mathbf{x}_{ij} - b_i] - 1 + \xi_{ij} \geqslant 0, \\
& \xi_{ij} \geqslant 0, \\
& i \in \{1, \ldots, P\}, \\
& j \in \{0, \ldots, D\}.
\end{aligned}
\tag{8}
$$

Введем обозначение:

$$
\hat{\mathbf{X}} = (\mathbf{A}^{-1})^{\mathrm{T}}\mathbf{X}. \tag{9}
$$

Найдём двойственную задачу:

$$\mathcal{L}(\mathbf{w}', \mathbf{b}, \xi, \lambda, r) = \frac{1}{2}\|\mathbf{w}'\|^2 + C\sum_{ij}\xi_{ij} - \sum_{ij}\lambda_{ij}\left(y_{ij}[\langle \mathbf{A}^{-1}\left(\mathbf{w}' + \mathbf{B}\right), \mathbf{x}_{ij}\rangle - b_i] - 1 + \xi_{ij}\right) - \sum_{ij} r_{ij}\xi_{ij}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}'} = \mathbf{w}' - \sum \lambda_{ij} y_{ij}\langle \mathbf{A}^{-1}\left(\mathbf{w}' + \mathbf{B}\right), \mathbf{x}_{ij}\rangle'_{w'} = 0 \rightarrow \mathbf{w}' = \sum_{ij}\lambda_{ij}y_{ij}\mathbf{A}^{-1}\mathbf{x}_{ij}$$

$$\forall i, \frac{\partial \mathcal{L}}{\partial b_i} = \sum_j \lambda_{ij}y_{ij} = 0$$

$$\forall (i,j), \frac{\partial \mathcal{L}}{\partial \xi_{ij}} = C - \lambda_{ij} - r_{ij} = 0 \rightarrow \lambda_{ij} + r_{ij} = C \tag{10}$$

$$\begin{aligned}\mathcal{L}(\lambda, r) &= \frac{1}{2}\langle \sum_{ij}\lambda_{ij}y_{ij}\mathbf{A}^{-1}\mathbf{x}_{ij}, \sum_{ij}\lambda_{ij}y_{ij}\mathbf{A}^{-1}\mathbf{x}_{ij}\rangle + C\sum_{ij}\xi_{ij} - \sum_{ij}\lambda_{ij}y_{ij}\langle \mathbf{A}^{-1}\sum_{ij}\lambda_{ij}y_{ij}\mathbf{A}^{-1}\mathbf{x}_{ij}, \mathbf{x}_{ij}\rangle - \\ &- \sum_{ij}\lambda_{ij}y_{ij}\langle \mathbf{A}^{-1}\mathbf{B}, \mathbf{x}_{ij}\rangle + \sum_{ij}\lambda_{ij}y_{ij}b_i + \sum_{ij}\lambda_{ij} - \sum_{ij}\lambda_{ij}\xi_{ij} - \sum_{ij} r_{ij}\xi_{ij} = \\ &= \frac{1}{2}\sum_{ij}\sum_{pq}\lambda_{ij}\lambda_{pq}y_{ij}y_{pq}\langle \mathbf{A}^{-1}\mathbf{x}_{ij}, \mathbf{A}^{-1}\mathbf{x}_{pq}\rangle + \sum_{ij}\xi_{ij}(C - \lambda_{ij} - r_{ij}) - \sum_{ij}\sum_{pq}\lambda_{ij}\lambda_{pq}y_{ij}y_{pq}\langle \mathbf{A}^{-2}\mathbf{x}_{ij}, \mathbf{x}_{pq}\rangle - \\ &- \sum_{ij}\lambda_{ij}y_{ij}\langle \mathbf{A}^{-1}\mathbf{B}, \mathbf{x}_{ij}\rangle + \sum_i b_i\sum_j \lambda_{ij}y_{ij} + \sum_{ij}\lambda_{ij}\end{aligned} \tag{11}$$

1. $\sum_{ij}\xi_{ij}(C - \lambda_{ij} - r_{ij}) = 0$ из ограничений;
2. $\sum_i b_i \sum_j \lambda_{ij}y_{ij} = 0$ из ограничений;
3. $\langle \mathbf{A}^{-2}\mathbf{x}, \mathbf{x}\rangle = \langle \mathbf{A}^{-1}\mathbf{x}, \mathbf{A}^{-1}\mathbf{x}\rangle$, т.к. $\mathbf{x}^\top\mathbf{A}^{-2}\mathbf{x} = \mathbf{x}^\top\mathbf{A}^{-1}\mathbf{A}^{-1}\mathbf{x} = (\mathbf{A}^{-1}\mathbf{x})^\top\mathbf{A}^{-1}\mathbf{x}$

   $\rightarrow$ введём обозначение: $\widehat{\mathbf{X}} = (\mathbf{A}^{-1})^{\mathrm{T}}\mathbf{X} = \mathbf{A}^{-1}\mathbf{X}$, т.к. $\mathbf{A}$ симметрична.

$$\begin{aligned}\mathcal{L}(\lambda) &= -\frac{1}{2}\sum_{ij}\sum_{pq}\lambda_{ij}\lambda_{pq}y_{ij}y_{pq}\langle \widehat{\mathbf{x}_{ij}}, \widehat{\mathbf{x}_{pq}}\rangle - \sum_{ij}\lambda_{ij}y_{ij}\langle \mathbf{A}^{-1}\mathbf{B}, \mathbf{x}_{ij}\rangle + \sum_{ij}\lambda_{ij} = \\ &= -\frac{1}{2}\sum_{ij}\sum_{pq}\lambda_{ij}\lambda_{pq}y_{ij}y_{pq}\langle \widehat{\mathbf{x}_{ij}}, \widehat{\mathbf{x}_{pq}}\rangle + \sum_{ij}\lambda_{ij}\left(1 - y_{ij}\langle \mathbf{B}, \widehat{\mathbf{x}_{ij}}\rangle\right)\end{aligned} \tag{12}$$

Двойственная задача: $\arg\max_{\lambda}\mathcal{L}(\lambda)$

Значит, исходная задача по теореме Каруша-Куна-Таккера эквивалентна двойственной:

$$\begin{aligned}\underset{\lambda_{ij}}{\text{minimize:}}\quad & \frac{1}{2}\sum_{(i,j),(p,q)}\lambda_{ij}\lambda_{pq}y_{ij}y_{pq}\langle \hat{\mathbf{x}}_{ij}, \hat{\mathbf{x}}_{pq}\rangle + \sum_{ij}\lambda_{ij}\left(y_{ij}\langle \mathbf{B}, \widehat{\mathbf{x}_{ij}}\rangle - 1\right) \\ \text{subject to:}\quad & 0 \leqslant \lambda_{ij} \leqslant C, \\ & \forall i, \sum_j \lambda_{ij}y_{ij} = 0 \\ & i \in \{1, \ldots, P\}, \\ & j \in \{0, \ldots, D\}.\end{aligned} \tag{13}$$

4 Эксперимент

5 Заключение

ЛитератураReferences

[1] Maria Kadukova, Sergei Grudinin. **Convex-PL: a novel knowledge-based potential for protein-ligand interactions deduced from structural databases using convex optimization**. Journal of Computer-Aided Molecular Design, October 2017, Volume 31, Issue 10, pp 943–958.

[2] Maria Kadukova and Sergei Grudinin. **Docking of small molecules to farnesoid X receptors using AutoDock Vina with the Convex-PL potential : lessons learned from D3R Grand Challenge 2**. J. Comput.-Aided Mol. Des., 2017.

[3] Sergei Grudinin, Maria Kadukova, Andreas Eisenbarth, Simon Marillet, Frédéric Cazals. **Predicting binding poses and affinities for protein-ligand complexes in the 2015 D3R Grand Challenge using a physical model with a statistical parameter estimation**. J Comput Aided Mol Des. 2016 Sep;30(9):791-804. Epub 2016 Oct 7.

[4] S.P. Boyd and L. Vandenberghe. **Convex optimization**. Cambridge Univ Press, 2004.