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

Илья Игашов, Сергей Грудинин, Мария Кадукова, В.В. Стрижов

Аннотация

В работе рассматривается задача оценки свободной энергии связывания белка с лигандом посредством оптимизациии скоринговой функции Convex-PL. Оптимизация сочетает в себе классификацию, основанную на методе опорных векторов, и регрессию, использующую различные функции потерь. Использование одной лишь задачи классификации для предсказания энергии связывания приводит к недостаточно высокой корреляции предсказаний с экспериментальными значениями, в то время как использование только регрессии приводит к переобучению. В этой работе предлагается посторить алгоритм, который объединяет классификацию и регрессию, решает данные проблемы и демонстрирует высокое качество оценки энергии связывания. Для проверки работы алгоритма будут использоваться данные, состоящие из комплексов белков и лигандов, для которых необходимо найти наилучшую позу лиганда или предсказать свободную энергию связывания. Также ниже будут представлены результаты работы алгоритма на датасете, состоящем из белков, для которых нужно найти наиболее сильно связывающий лиганд. Результаты будут получены как на выборках пониженной размерности, так и на полных выборках с большой размерностью.

1. Введение

Развитие вычислительных методов и появление новых подходов в молекулярном моделировании дает широкие возможности в области изучения химических соединений. В частности, популярный ныне метод виртуального скрининга применяется при разработке новых лекарственных препаратов для поиска и анализа химических соединений, обладающих необходимыми биологическими свойствами [1].

Среди методов молекулярного моделирования широко распространен молекулярный докинг, позволяющий предсказать взаимную ориентацию молекул, наиболее выгодную для образования устойчивого комплекса [2]. Поскольку для решения данной задачи требуется анализ и обработка огромного количества данных, содержащих в себе информацию о химических комплексах и их характеристиках, она является вычислительно трудоемкой и требует решений, обладающих высокой производительностью.

Образование комплекса «белок-лиганд» можно рассматривать как термодинамическое событие, описываемое постоянной степени сродства (аффинности связывания)

соединения, которая напрямую сопряжена со свободной энергией связывания белка с лигандом. Нативные конформации отвечают минимуму энергии связывания. Задача молекулярного докинга — предсказать эти взаимные расположения и отвечающие им энергетические значения. Свободная энергия связывания зависит от множества факторов, включающих в себя не только взаимодействие белка с лигандом, но также сольватацию и энтропийные факторы. Строгий подсчет значений энергии связывания потребовал бы семплирования всего конфигурационного пространства, что является, с точки зрения вычислений, крайне затратной задачей ввиду высокой размерности пространства [3]. В последние годы для решения данной задачи был предложен ряд аппроксимирующих алгоритмов, оценивающих значение энергии связывания на основе скоринговых функций [4].

Одной из таких функций является Convex-PL [3]. Данная функция зависит от взаимного расположения белков и лигандов и достигает минимума на нативной конформации бинарной системы. Основная идея построения Convex-PL состоит в декомпозиции молекул белков и лигандов на отдельные атомы из некоторого конечного набора и последующем подсчете значения функционала на всевозможных комбинациях различных пар атомов, учитывая априорные распределения плотностей расстояний между ними. Как показано [3], данную функцию можно приблизить полиномом конечной степени с так называемыми «структурными» коэффициентами, характеризующими взаимное расположение белков и лигандов, и найти эти коэффициенты разложения методами выпуклой оптимизации [5].

Для решения описанной выше оптимизационной задачи были предложены два подхода. В работе [3] описывается алгоритм классификации, основанный на методе опорных векторов [6]. Данный алгоритм был протестирован с помощью D3R Grand Challenge [7] и CASF 2013. Полученные на CASF 2013 результаты по предсказанию нативных и около-нативных конформаций превзошли показатели других 20 методов, протестированных на тех же данных ранее. В работе [8] был предложен регрессионный алгоритм, использующий linear ridge regerssion (LRR) и kernel ridge regression (KRR) [9, 10, 11]. Данный метод был протестирован на датасетах MAP4K4 и HSP90 и продемонстрировал хорошие результаты [8]. Однако, как показали исследования [3, 8], оба подхода также имеют и слабые стороны: использование одной лишь задачи классификации для предсказания энергии связывания приводит к недостаточно высокой корреляции предсказаний с экспериментальными значениями, в то время как использование только регрессии приводит к переобучению. Кроме того, исследования, представленные в [3, 8], были проведены на данных пониженной размерности. В этой работе представлен алгоритм молекулярного докинга для предсказания нативных конформаций комплексов «белок-лиганд» и значений энергии связывания данных соединений. Описанный ниже алгоритм решает задачу поиска минимума скоринговой функции Convex-PL методами выпуклой оптимизации, сочетая в себе задачи классификации и регрессии.

2. Модель взаимодействий

Пусть имеется P нативных комплексов белков-лигандов $\{C_{i0}\}_{i=1}^{P}$. Применив к лигандам изометрические преобразования, сгенерируем для каждой нативной позы D ненативных поз $\{C_{ij}\}_{j=1}^{D}$. Таким образом, для каждого из P комплексов имеем $(D+1)$ конформаций: одну нативную и D ненативных. Требуется найти скоринговый функционал E, который удовлетворяет следующим неравенствам:

$$E(C_{i0}) < E(C_{ij}) \ \forall i \in \{1, \ldots, P\}, \ \forall j \in \{1, \ldots, D\}. \tag{1}$$

В качестве такого функционала будем рассмаривать свободную энергию связывания белков с лигандами, определенную для всех возможных комплексов. Для упрощения формы функционала сделаем ряд допущений.

Во-первых, будем рассматривать комплекс "белок-лиганд" как набор атомов, каждый из которых имеет некоторый тип. Тип каждого атома зависит от его химических свойств, таких как номер элемента в периодической таблице, аромат, гибридизация, полярность и т.д. Пусть M_1 – количество типов атомов лиганда, а M_2 – количество типов атомов белка. Тогда получим всего $M_1 \times M_2$ различых атомных взаимодействий.

Во-вторых, будем считать, что E определяется только взаимодействиями между парами атомов рассматриваемого комплекса. При этом в каждой паре первый атом является атомом лигнда, а второй – атомом белка. Кроме того, будем рассматривать только те пары, в которых расстояние между атомами не превышает некоторой пороговой величины $r_{\max}$. В качестве $r_{\max}$ возьмем значение 10Å, как это было сделано в других работах [12, 13, 14, 15, 16, 17] ранее.

В-третьих, будем считать, что E зависит только от распределения расстояний между взаимодействующими атомами.

И, наконец, предположим, что E является линейным функционалом и имеет вид:

$$E(n(r)) = \sum_{k=1}^{M_1} \sum_{l=1}^{M_2} \int_0^{r_{max}} n^{kl}(r) f^{kl}(r) dr, \tag{2}$$

где $f^{kl}(r)$ – неизвестные функции взаимодествия между атомами типов k и l. Будем называть их *скоринговыми потенциалами*. Функции $n^{kl}(r)$ – численные плотности распределений пар атомов типов k и l по расстоянию r между ними:

$$n^{kl}(r) = \sum_{i,j} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(r - r_{ij})^2}{2\sigma^2}\right], \tag{3}$$

где σ^2 – стандартное отклонение (константа). Сумма берется по всем парам (i, j) атомов с типами k и l соответственно, у которых расстояние между атомами не превышает пороговой величины $r_{\max}$, атом i принадлежит лиганду, а атом j – белку.

Разложим неизвестные скоринговые потенциалы $f^{kl}(r)$ и плотности $n^{kl}(r)$ по полиномиальному базису:

$$\begin{aligned} f^{kl}(r) &= \sum_q w_q^{kl}\psi_q(r), \\ n^{kl}(r) &= \sum_q x_q^{kl}\psi_q(r), \end{aligned} \tag{4}$$

где $\psi_q(r)$ – ортогональные базисные функции на интервале $[0, r_{\max}]$, а w_q^{kl} и x_q^{kl} – коэффициенты разложения функций $f^{kl}(r)$ и $n^{kl}(r)$ соответственно. Поскольку базисные функции ортогональны, справедливо следующее соотношение:

$$\int_0^{r_{\max}} \psi_i(r)\psi_j(r)\Omega(r)dr = \delta_{ij},\ r \in [0, r_{\max}], \tag{5}$$

где $\Omega(r)$ – некоторая неотрицательная весовая функция на $[0, r_{\max}]$, δ_{ij} – символ Кронекера. Из данного условия (5) ортогональности базисных функций могут быть найдены коэффициенты раложения w_q^{kl} и x_q^{kl}:

$$\begin{aligned} w_q^{kl} &= \int_0^{r_{\max}} f_{kl}(r)\psi_q(r)dr, \\ x_q^{kl} &= \int_0^{r_{\max}} n_{kl}(r)\psi_q(r)dr, \end{aligned} \tag{6}$$

Таким образом, функционал E можно записать в следующем виде:

$$E(n(r)) = \sum_{k=1}^{M_1}\sum_{l=1}^{M_2}\sum_{pq}^{\infty} w_q^{kl}x_p^{kl}\int_0^{r_{\max}} \psi_q(r)\psi_p(r)\Omega(r)dr. \tag{7}$$

Учитывая условие ортогональности (5) и ограничиваясь порядком разложения Q, получаем приближенное выражение скорингового функционала:

$$\begin{gathered} E(n(r)) \approx \sum_{k=1}^{M_1}\sum_{l=1}^{M_2}\sum_{q=0}^{Q} w_q^{kl}x_q^{kl} = \langle \mathbf{w}, \mathbf{x}\rangle, \\ \mathbf{w}, \mathbf{x} \in \mathbb{R}^{Q\times M_1\times M_2}. \end{gathered} \tag{8}$$

Неизвестный вектор $\mathbf{w}$ будем называть *скоринговым вектором*, а вектор $\mathbf{x}$ – *структурным вектором*. Структурный вектор известен для решения задачи: его можно получить из данных. Для оценки энергии связывания в работе [8] использовался порядок разложения $Q = 10$, и рассматривались $M_1 = 48$ типов атомов.

3. Постановка задачи

Пусть имеется P нативных соединений белков и лигандов, где i-е соединение описывается структурным вектором $\mathbf{x}_{i0}^{\text{nat}}$, и $(P \times D)$ ненативных соединений с соответствующими структурными векторами $\mathbf{x}_{ij}^{\text{nonnat}}$. Учитывая (8) и тот факт, что минимум

энергии связывания соответствует нативным конформациям соединений белков и лигандов, справедливы следующие неравенства:

$$\begin{aligned}
&\langle \mathbf{x}_{i0}^{\text{nat}}, \mathbf{w} \rangle < \langle \mathbf{x}_{ij}^{\text{nonnat}}, \mathbf{w} \rangle, \\
&\langle \mathbf{x}_{ij}^{\text{nonnat}} - \mathbf{x}_{i0}^{\text{nat}}, \mathbf{w} \rangle > 0, \\
&i = 1, \ldots, P, \\
&j = 1, \ldots, D.
\end{aligned} \tag{9}$$

В такой постановке задачи требуется отыскать неизвестный скоринговый вектор $\mathbf{w}$, который бы решал задачу классификации, а именно, определял бы, является ли конформация белка и лиганда со структурным вектором $\mathbf{x}$ нативной или ненативной.

Система неравенств (9) может иметь ноль, одно или бесконечное число решений [3, 6]. Чтобы получить единственное решение задачи (9), переформулируем ее в виде задачи квадратичной оптимизации с мягким зазором [5]:

$$\begin{aligned}
\underset{\mathbf{w}, b_i, \xi_{ij}}{\text{minimize:}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{ij} \xi_{ij} \\
\text{subject to:} \quad & y_{ij}[\langle \mathbf{w}, \mathbf{x}_{ij} \rangle - b_i] - 1 + \xi_{ij} \geq 0, \\
& \xi_{ij} \geq 0, \\
& i \in \{1, \ldots, P\}, \\
& j \in \{0, \ldots, D\},
\end{aligned} \tag{10}$$

где скоринговый вектор $\mathbf{w}$, вектор смещений b_i и переменные невязки ξ_{ij} – оптимизируемые параметры модели, y_{ij} – класс j-й позы i-го соединения ($y_{i0} = 1$ для нативной позы и $y_{ij} = -1,\ j \in \{1, \ldots, D\}$, для ненативной позы), C – некоторый коэффициент регуляризации. Таким образом, можно получить классификатор нативных и ненативных поз соединения, решив оптимизационную задачу (10).

Кроме того, для предсказания свободной энергии связывания белка с лигандом можно решать задачу регрессии [8]:

$$\begin{aligned}
\underset{\mathbf{w}}{\text{minimize:}} \quad & \sum_i [\langle \mathbf{w}, \mathbf{x}_{i0} \rangle - s_i]^2 + \alpha \|\mathbf{w}\|^2, \\
& i \in \{1, \ldots, P\},
\end{aligned} \tag{11}$$

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

Оба описанных выше подхода в конечном итоге решают одну и ту же задачу поиска неизвестного скорингового вектора $\mathbf{w}$. Объединив эти методы, мы также будем решать задачу оптимизации модели, предсказывающей энергию связывания белков с

лигандами:

$$\begin{aligned} \underset{\mathbf{w},b_i,\xi_{ij}}{\text{minimize:}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{ij}\xi_{ij} + C_r\sum_i f(\mathbf{x}_{i0},\mathbf{w},s_i) \\ \text{subject to:} \quad & y_{ij}[\langle \mathbf{w},\mathbf{x}_{ij}\rangle - b_i] - 1 + \xi_{ij} \geq 0, \\ & \xi_{ij} \geq 0, \\ & i \in \{1,\ldots,P\}, \\ & j \in \{0,\ldots,D\}, \end{aligned} \tag{12}$$

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

Поднимем размерность пространства векторов $\mathbf{w}$ и $\mathbf{x}$, избавившись от вектора смещения b_i:

$$\begin{aligned} & \mathbf{w}^{\mathrm{T}} \leftarrow (\mathbf{w}^{\mathrm{T}}, b_1, \ldots, b_P), \\ & \mathbf{x}_{1j}^{\mathrm{T}} \leftarrow (\mathbf{x}_{1j}^{\mathrm{T}}, -1, 0, \ldots, 0), \\ & \mathbf{x}_{2j}^{\mathrm{T}} \leftarrow (\mathbf{x}_{2j}^{\mathrm{T}}, 0, -1, 0, \ldots, 0), \\ & \ldots \\ & \mathbf{x}_{Pj}^{\mathrm{T}} \leftarrow (\mathbf{x}_{Pj}^{\mathrm{T}}, 0, \ldots, 0, -1), \\ & j \in \{0,\ldots,D\}. \end{aligned} \tag{13}$$

Теперь $\mathbf{w}, \mathbf{x}_{ij} \in \mathbb{R}^{(Q\times M_1\times M_2)+P}$, а задача оптимизации (12) принимает вид:

$$\begin{aligned} \underset{\mathbf{w},\xi_{ij}}{\text{minimize:}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{ij}\xi_{ij} + C_r\sum_i f(\mathbf{x}_{i0},\mathbf{w},s_i) \\ \text{subject to:} \quad & y_{ij}\langle \mathbf{w},\mathbf{x}_{ij}\rangle - 1 + \xi_{ij} \geq 0, \\ & \xi_{ij} \geq 0, \\ & i \in \{1,\ldots,P\}, \\ & j \in \{0,\ldots,D\}. \end{aligned} \tag{14}$$

С помощью замены переменных сведем два квадратичных слагаемых в целевой функции из задачи (14) к одному. Функция потерь регрессии Mean Squared Error выражается формулой:

$$f(\mathbf{x}_{i0},\mathbf{w},s_i) = (\mathbf{w}^{\mathrm{T}}\mathbf{x}_{i0} - s_i)^2. \tag{15}$$

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

(14) принимают вид:

$$\frac{1}{2}\|\mathbf{w}\|^2 + C_r \sum_i f(\mathbf{x}_{i0}, \mathbf{w}, s_i) = \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + C_r(\mathbf{w}^{\mathrm{T}}\mathbf{X} - \mathbf{s})^2 =$$

$$\frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + C_r\mathbf{w}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{w} - 2C_r\mathbf{w}^{\mathrm{T}}\mathbf{X}\mathbf{s} + C_r\mathbf{s}^{\mathrm{T}}\mathbf{s} =$$

$$\mathbf{w}^{\mathrm{T}}\left(\frac{1}{2}\mathbf{I} + C_r\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)\mathbf{w} - 2C_r\mathbf{w}^{\mathrm{T}}\mathbf{X}\mathbf{s} + C_r\mathbf{s}^{\mathrm{T}}\mathbf{s} =$$

$$\left(\left[\frac{1}{2}\mathbf{I} + C_r\mathbf{X}\mathbf{X}^{\mathrm{T}}\right]^{\frac{1}{2}}\mathbf{w}\right)^2 - 2C_r\left(\mathbf{w}^{\mathrm{T}}\left[\frac{1}{2}\mathbf{I} + C_r\mathbf{X}\mathbf{X}^{\mathrm{T}}\right]^{\frac{1}{2}}\right)\left(\left[\frac{1}{2}\mathbf{I} + C_r\mathbf{X}\mathbf{X}^{\mathrm{T}}\right]^{-\frac{1}{2}}\mathbf{X}\mathbf{s}\right) =$$

$$\left(\left[\frac{1}{2}\mathbf{I} + C_r\mathbf{X}\mathbf{X}^{\mathrm{T}}\right]^{\frac{1}{2}}\mathbf{w} - C_r\left[\frac{1}{2}\mathbf{I} + C_r\mathbf{X}\mathbf{X}^{\mathrm{T}}\right]^{-\frac{1}{2}}\mathbf{X}\mathbf{s}\right)^2 + C_r\mathbf{s}^{\mathrm{T}}\mathbf{s} - C_r^2\left[\frac{1}{2}\mathbf{I} + C_r\mathbf{X}\mathbf{X}^{\mathrm{T}}\right]^{-1}\mathbf{X}\mathbf{s}.$$

Введем замену переменных:

$$\begin{aligned}
&\mathbf{w}' = \mathbf{A}\mathbf{w} - \mathbf{B}, \text{ где} \\
&\mathbf{A} = \left[\frac{1}{2}\mathbf{I} + C_r\mathbf{X}\mathbf{X}^{\mathrm{T}}\right]^{\frac{1}{2}}, \\
&\mathbf{B} = C_r\left[\frac{1}{2}\mathbf{I} + C_r\mathbf{X}\mathbf{X}^{\mathrm{T}}\right]^{-\frac{1}{2}}\mathbf{X}\mathbf{s}.
\end{aligned} \tag{16}$$

Тогда, учитывая, что

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

задача оптимизации принимает вид:

$$\begin{aligned}
\underset{\mathbf{w}', b_i, \xi_{ij}}{\text{minimize:}} \quad & \frac{1}{2}\|\mathbf{w}'\|^2 + C\sum_{ij}\xi_{ij} \\
\text{subject to:} \quad & y_{ij}[\left(\mathbf{A}^{-1}\left(\mathbf{w}' + \mathbf{B}\right)\right)^{\mathrm{T}}\mathbf{x}_{ij} - b_i] - 1 + \xi_{ij} \geq 0, \\
& \xi_{ij} \geq 0, \\
& i \in \{1, \ldots, P\}, \\
& j \in \{0, \ldots, D\}.
\end{aligned} \tag{17}$$

Введем обозначение:

$$\hat{\mathbf{X}} = (\mathbf{A}^{-1})^{\mathrm{T}}\mathbf{X}. \tag{18}$$

По теореме Каруша-Куна-Таккера [5], задача (17) эквивалентна двойственной задаче:

$$\begin{aligned} \underset{\lambda_{ij}}{\text{minimize:}} \quad & -\sum_{ij} \lambda_{ij} + \frac{1}{2} \sum_{(i,j),(p,q)} \lambda_{ij}\lambda_{pq} y_{ij} y_{pq} \langle \hat{\mathbf{x}}_{ij}, \hat{\mathbf{x}}_{pq} \rangle \\ \text{subject to:} \quad & 0 \geq \lambda_{ij} \geq C, \\ & i \in \{1, \ldots, P\}, \\ & j \in \{0, \ldots, D\}. \end{aligned} \tag{19}$$

4. Эксперимент

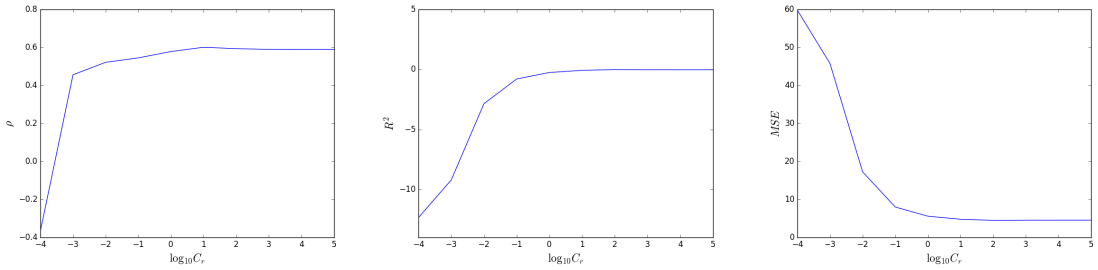
Для обучения модели были выбраны данные из базы PDBBind [18, 19], которая содержит экспериментально определенные комплексы белков и лигандов и соответствующие им измеренные значения постоянной аффинности связывания. Создатели PDBBind предлагают для исследований два датасета. Собранный вручную "general set" содержит 14620 комплексов белков и лигандов, а также значения аффинности связывания (K_d, K_i и IC_{50}). Кроме того, авторы отдельно отобрали данные лучшего качества и собрали дополнительный "refined set" , также состоящий из комплексов белков и лигандов и соответствующих им значений аффинности связывания.

В представленном эксперименте используются данные из "refined set" , для которых известно значение константы K_i: всего 39444 конформации, 2076 из которых являются нативными, и 37368 – ненативными. В данном наборе каждому соединению соответствует одна нативная поза с известным значением константы K_i и 18 ненативных поз, для которых значение K_i неизвестно. Таким образом, $P = 2076$ и $D = 18$. В базе PDBBind рассматривается $M_1 = 40$ типов атомов лигандов и $M_2 = 23$ типов атомов белков, а значение Q принято равным 7. Таким образом, каждая конформация характеризуется структурным вектором $\mathbf{x}_{ij}$ размерности $Q \times M_1 \times M_2 = 6440$.

Из исходной выборки `X` и соответствующего вектора классов конформаций `y` на обучающую выборку (`X_train, y_train`) было выделено 60% объектов, а на тестовую – все нативные комплексы и соответствующие значения постоянной аффинности связывания (`X_test, s_test`) из оставшихся 40% выборки `X`. Для замены переменных из обучающей выборки были выделены исключительно нативные конформации с известными значениями постоянной K_i, которые образовали выборку `X_nat_train` и целевой вектор `s_train` соответственно. С помощью этих данных были вычислены матрицы $\mathbf{A}$ и $\mathbf{B}$.

Чтение и обработка данных, включая замену переменных и вычисление ошибки предсказаний, проводились на языке `Python` с использованием пакетов `NumPy, SciPy` и `scikit-learn`. Поиск минимума функции (14) и оптимизация скорингового вектора $\mathbf{w}'$ осуществлялись с помощью библиотеки `LIBLINEAR` [20], написанной на языке `C++`. В данной библиотеке реализован метод L2-regularized L1-loss Support Vector Classification (Solving Dual) [21], который решает задачу (17) оптимизации вектора $\mathbf{w}'$ с помощью перехода в двойственное пространство и решения задачи (19). Для оценки качества модели измерялись коэффициент корреляции Пирсона ρ, коэффициент детерминации R^2 и среднеквадратичное отклонение MSE.

Для подбора оптимального значения коэффициента регуляризации C_r, который использовался в замене переменных (16), алгоритм был запущен на сетке значений $\{0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000, 10000, 100000\}$. Наилучшие результаты наблюдаются при значении $C_r = 100$ (Рис. 1).



(a) (b) (c)

Рис. 1: Зависимость (a) коэффициента корреляции Пирсона ρ, (b) коэффициента детерминации R^2, (c) среднеквадратичной ошибки MSE от значения логарифма коэффициента регуляризации C_r.

Для значения $C_r = 100$ и $C = 20000$ были проведены тесты на классификацию поз новых соединений. Для этого использовался докинг-тест из бенчмарка CASF, в котором предсказывается правильность угадывания top 1,2,3 поз с RMSD $< 1, 2, 3$ по сравнению с нативной позой для 195 комплексов белков и лигандов, для каждого из которых авторами теста сгенерировано около 100 поз с RMSD ≤ 10.0. В отличие от декоев из PDBBind, которые сгенерированы поворотами и трансляцией, в этом тесте декои сгенерированы более сложными докинг-алгоритмами. В скоринг-тесте считается коэффициент корреляции Пирсона между предсказанными скорами 195 комплексов (в нативных конформациях) и известными $logK_i$ или K_d. Полученные результаты приведены в Таблице 1 (t1 native − процент комплексов, для которых угаданная поза является нативной. t3q3 − процент комплексов, для которых первые три позы с наилучшим скором оказались в пределах RMSD=3 Ангстрема по сравнению с нативной позой).

t1 native 28.2	t1q1 64.1	t1q2 72.3	t1q3 77.4	t1q1n 55.9	t1q2n 67.2	t1q3n 73.3
t2 native 46.7	t2q1 78.5	t2q2 84.1	t2q3 88.2	t2q1n 68.7	t2q2n 78.5	t2q3n 85.1
t3 native 56.9	t3q1 83.1	t3q2 88.7	t3q3 92.3	t3q1n 73.3	t3q2n 84.1	t3q3n 90.3

Таблица 1: Результаты на докинг-тесте из бенчмарка CASF

В данном тесте коэффициент корреляции Пирсона составил $\rho = 0.521182$.

5. Заключение

Полученная модель показала достойные результаты на докинг-тестах из бенчмарка CASF, однако результаты могут быть выше, если точнее подобрать коэффициенты

регуляризации и работать с данными повышенной размерности. Предлагается расширить признаковое описание комплексов, добавив дополнительные водные дескрипторы, а также признаки, определяющие количества связей в комплексе и их вращения.

Список литературы

[1] Christoph Sotriffer and Hans Matter. **Virtual Screening: Principles, Challenges, and Practical Guidelines**. Wiley Online Library, 10.1002/9783527633326.ch7 edition, 2011.

[2] Lengauer T, Rarey M (Jun 1996). **"Computational methods for biomolecular docking". Current Opinion in Structural Biology**. 6 (3): 402–6.

[3] 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.

[4] Xuan-Yu Meng, Hong-Xing Zhang, Mihaly Mezei, and Meng Cui. **Molecular Docking: A powerful approach for structure-based drug discovery**. Curr Comput Aided Drug Des. 2011 Jun 1; 7(2): 146–157.

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

[6] V. Vapnik. **The nature of statistical learning theory**. Springer, 2000.

[7] 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.

[8] 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.

[9] D. Barber. **Bayesian reasoning and machine learning**. Cambridge University Press, Cambridge, 2012.

[10] Kevin Vu, John Snyder, Li Li, Matthias Rupp, Brandon F. Chen, Tarek Khelif, Klaus-Robert Müller, Kieron Burke. **Understanding Kernel Ridge Regression: Common behaviors from simple functions to density functionals**.

[11] Press, WH; Teukolsky, SA; Vetterling, WT; Flannery, BP (2007). **"Section 19.5. Linear Regularization Methods". Numerical Recipes: The Art of Scientific Computing (3rd ed.)**. New York: Cambridge University Press. ISBN 978-0-521-88068-8.

[12] Sheng-You Huang and Xiaoqin Zou. **An iterative knowledge-based scoring function for protein–protein recognition**. Proteins: Struct., Funct., Bioinf., 72(2):557–579, 2008.

[13] Gwo-Yu Chuang, Dima Kozakov, Ryan Brenke, Stephen R Comeau, and Sandor Vajda. **Dars (decoys as the reference state) potentials for protein-protein docking**. Biophys. J., 95(9):4217–4227, 2008.

[14] Vladimir N Maiorov and Gordon M Grippen. **Contact potential that recognizes the correct folding of globular proteins**. J. Mol. Biol., 227(3):876–888, 1992.

[15] Jian Qiu and Ron Elber. **Atomically detailed potentials to recognize native and approximate protein structures**. Proteins: Struct., Funct., Bioinf., 61(1):44–55, 2005.

[16] Dror Tobi and Ivet Bahar. **Optimal design of protein docking potentials: Efficiency and limitations**. Proteins: Struct., Funct., Bioinf., 62(4):970–981, 2006.

[17] Myong-Ho Chae, Florian Krull, Stephan Lorenzen, and Ernst-Walter Knapp. **Predicting protein complex geometries with a neural network**. Proteins: Struct., Funct., Bioinf., 78(4):1026–1039, 2010.

[18] Renxiao Wang, Xueliang Fang, Yipin Lu, Chao-Yie Yang, and Shaomeng Wang. **The PDBbind Database: Methodologies And Updates**. J. Med. Chem., 48(12):4111–9, June 2005.

[19] Renxiao Wang, Xueliang Fang, Yipin Lu, and Shaomeng Wang. **The PDBbind Database:**

Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures. J. Med. Chem., 47(12):2977–80, June 2004.

[20] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, Chih-Jen Lin. **LIBLINEAR: A Library for Large Linear Classification**. The Journal of Machine Learning Research archive, Volume 9, 6/1/2008, Pages 1871-1874.

[21] Cho-Jui Hsieh, Kai-Wei Chang, Chih-Jen Lin, S. Sathiya Keerthi, and Sellamanickam Sundararajan. **A dual coordinate descent method for large-scale linear SVM**. In Proceedings of the Twenty Fifth International Conference on Machine Learning (ICML), 2008.