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

Грачева А. С., Соавтор И. О., Фамилия И. О.
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. Можно генерировать несколько возможных положений лиганда 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 если использовать экспериментальные данные об аффинностях, то можно улучшить качество классификации. В эксперименте, описанном в данной статье, мы проверим эту гипотезу, а также постараемся решать оптимизационную задачу максимально эффективно вычислительно, чтобы иметь возможность 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:

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

*Работа выполнена при финансовой поддержке РФФИ, проект №00-00-00000. Научный руководитель: Стрижов В. В. Задачу поставил: Эксперт И. О. Консультант: Консультант И. О.

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.

2.1 Название параграфа.

Нет ограничений на количество разделов и параграфов в статье. Разделы и параграфы не нумеруются.

2.2 Теоретическую часть работы

желательно структурировать с помощью окружений Def, Axiom, Hypothesis, Problem, Lemma, Theorem, Corollary, State, Example, Remark.

ОпределениеDefinition 1. *Математический текст хорошо структурирован, если в нём выделены определения, теоремы, утверждения, примеры, и т. д., а неформальные рассуждения (мотивации, интерпретации) вынесены в отдельные параграфы.*

УтверждениеStatement 1. *Мотивации и интерпретации наиболее важны для понимания сути работы.*

ТеоремаTheorem 1. *Не менее* 90% *коллег, заинтересовавшихся Вашей статьёй, прочитают в ней не более* 10% *текста.*

Доказательство. Причём это будут именно те разделы, которые не содержат формул. ■

ЗамечаниеRemark 1. Выше показано применение окружений Def, Theorem, State, Remark, Proof.

3 Некоторые формулы

Образец формулы: $f(x_i, \alpha^\gamma)$.

Образец выключной формулы без номера:

$$y(x,\alpha) = \begin{cases} -1, & \text{если } f(x,\alpha) < 0; \\ +1, & \text{если } f(x,\alpha) \geqslant 0. \end{cases}$$

Образец выключной формулы с номером:

$$y(x,\alpha) = \begin{cases} -1, & \text{если } f(x,\alpha) < 0; \\ +1, & \text{если } f(x,\alpha) \geqslant 0. \end{cases} \tag{6}$$

Образец выключной формулы, разбитой на две строки с помощью окружения align:

$$R'_N(F) = \frac{1}{N}\sum_{i=1}^{N}\Big(P(+1\,|\,x_i)C\big(+1, F(x_i)\big) + \\ + P(-1\,|\,x_i)C\big(-1, F(x_i)\big)\Big). \tag{7}$$

Образцы ссылок: формулы (6) и (7).

4 Пример илюстрации

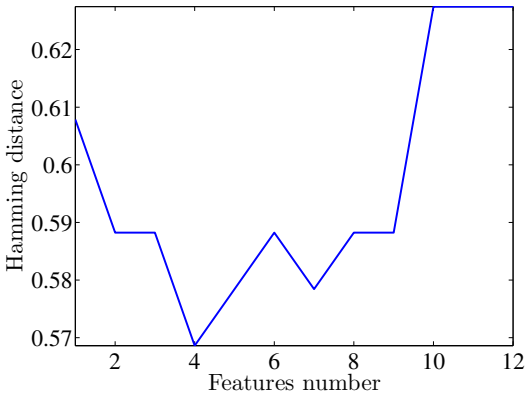
Рисунки вставляются командой `\includegraphics`, желательно с выравниванием по ширине колонки: `[width=\linewidth]`.

Практически все популярные пакеты рисуют графики с подписями, которые трудно читать на бумаге и на слайдах из-за малого размера шрифта. Шрифт на графиках (подписи осей и цифры на осях) должны быть такого же размера, что и основной текст.

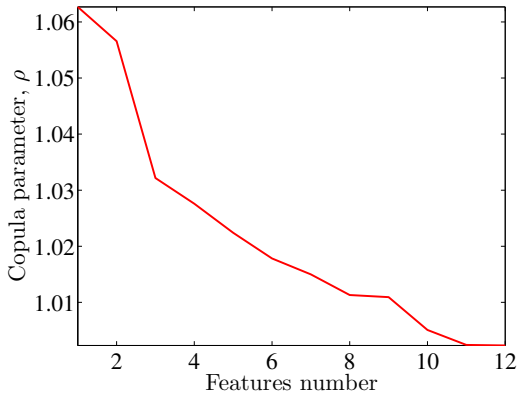
При значительном количестве рисунков рекомендуется группировать их в одном окружении `{figure}`, как это сделано на рис. 1.

5 Пример таблицы

Подпись делается *над таблицей*, см. таблицу 1.



(*а*) Первый рисунок



(*б*) Второй рисунок

Рис.Figure 1 Подпись должна размещаться под рисунком.

ТаблицаTable 1 Подпись размещается над таблицей.

Задача	CCEL	boosting
`Cancer`	$\mathbf{3.46} \pm 0.37$ (3.16)	4.14 ± 1.48
`German`	$\mathbf{25.78} \pm 0.65$ (1.74)	29.48 ± 0.93
`Hepatitis`	18.38 ± 1.43 (2.87)	19.90 ± 1.80

6 Заключение

Желательно, чтобы этот раздел был, причём он не должен дословно повторять аннотацию. Обычно здесь отмечают, каких результатов удалось добиться, какие проблемы остались открытыми.

ЛитератураReferences

ЛитератураReferences

[1] *Maria Kadukova, Sergei Grudinin* Maria Kadukova, Sergei Grudinin Convex-PL: a novel knowledge-based potential for protein-ligand interactions deduced from structural databases using convex optimization*Convex-PL: a novel knowledge-based potential for protein-ligand interactions deduced from structural databases using convex optimization* // Journal of ComputerAided Molecular Design, Springer Verlag *Journal of ComputerAided Molecular Design, Springer Verlag*, 2017, 31 (10), pp.943-958.